01010
01010
01010
information

Special Issue Reprint

Digital Privacy and Security, 2nd Edition

Edited by
José Braga de Vasconcelos, Hugo Barbosa and Carla Cordeiro

mdpi.com/journal/information

MDPI

Digital Privacy and Security, 2nd Edition

Digital Privacy and Security, 2nd Edition

Guest Editors

José Braga de Vasconcelos
Hugo Barbosa
Carla Cordeiro

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors

José Braga de Vasconcelos
Department of Computer Engineering and Information Systems (DEISI)
Lusófona University
Porto
Portugal

Hugo Barbosa
Department of Computer Engineering and Information Systems (DEISI)
Lusófona University
Porto
Portugal

Carla Cordeiro
Department of Computer Engineering and Information Systems (DEISI)
Lusófona University
Porto
Portugal

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Information* (ISSN 2078-2489), freely accessible at: www.mdpi.com/journal/information/special_issues/NC85V913K6.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-7468-2 (Hbk)
ISBN 978-3-7258-7469-9 (PDF)
https://doi.org/10.3390/books978-3-7258-7469-9

Contents

About the Editors

José Braga de Vasconcelos

José Braga de Vasconcelos is an Associate Professor at the University Lusófona, Portugal. PhD in Computer Science at the University of York, UK (2002). Research development in Knowledge Management and Engineering. Head of the Department of Computer Engineering and Information Systems at the University Lusófona, Porto. Professor of Algorithms and Data Structures and Software Engineering. Research integrated member of the Cognitive and People-Centric Computing (COPELABS Research Unit). Academic director at the University Lusófona of the Sustainable Information Technologies for Societies European Project which implements the Joint Master Degree (Erasmus Mundus) in Artificial Intelligence for Sustainable Societies (AISS).

Hugo Barbosa

Hugo Barbosa has been teaching Informatics Engineering at the Lusófona University—Porto University Center since 2010, and more recently at Porto Polytechnic (IPP) in the field of informatics. He collaborates with the Social Innovation and Interactive Systems (SIIS), an I&D by Porto Polytechnic. Currently, he is a PhD. candidate in Informatics Engineering at the Faculty of Engineering of Porto (FEUP), where he deepens his studies and research in serious games, virtual reality, simulation, player adaptability, as well as essential topics in cybersecurity and computer networks. He is also the Senior Member of the Portuguese Engineers Association (OE).

Carla Cordeiro

Carla Cordeiro holds a bachelor's and a master's degree in Electrical and Computer Engineering from the Faculty of Engineering at the University of Porto (FEUP). With nearly thirty years of experience in university teaching, she is a professor at Universidade Lusófona (Porto University Center) and serves as a guest professor at the Instituto Superior de Engenharia do Porto (ISEP). Her main areas of interest include data science, databases, decision support systems, control systems, nonlinear systems, and mathematics, as well as innovative teaching methodologies.

Preface

Since the publication of the first edition of 'Digital Privacy and Security', the digital landscape has evolved substantially. Emerging technologies, the proliferation of connected devices, the exponential growth of personal data, and the consolidation of generative artificial intelligence (AI) demand a constant re-evaluation of digital privacy and security strategies. This second edition expands the discussion by integrating recent research that reflects the current complexities of this rapidly changing environment and digital transformation acceleration for individuals and organizations.

The resilience of embedded systems against digital threats is a central theme in Optimized Digital Watermarking for Robust Information Security in Embedded Systems, where digital watermarking is used not only for protection but also as a mechanism of resilience against unauthorized manipulation. Security is now built into devices from the ground up, rather than added as an afterthought.

Federated learning, a promising tool for collaborative data analysis without centralization, brings specific privacy challenges. A Comparative Study of Privacy-Preserving Techniques in Federated Learning offers a comparative overview of the most effective approaches, showing the delicate balance required between performance and data protection.

In the realm of wearable devices, the continuous stream of data poses new vulnerabilities. Protecting Infinite Data Streams from Wearable Devices with Local Differential Privacy Techniques proposes local differential privacy solutions, demonstrating that even in scenarios of constant data collection, it is possible to safeguard user identity without compromising functionality.

The social dimension of cybersecurity gains relevance in Smartphone Privacy and Cyber Safety among Australian Adolescents, which reveals important gender differences in perception and behavior. Understanding these factors is key to developing effective educational policies and protection tools.

Cryptographic solutions have also advanced. Efficient Cryptographic Solutions for Unbalanced Private Set Intersection in Mobile Communication addresses mobile communication scenarios where shared data sets are uneven, proposing secure and efficient methods to preserve privacy even in resource-constrained environments.

In the corporate sphere, secure data sharing between businesses is vital in interdependent digital ecosystems. Privacy and Security Mechanisms for B2B Data Sharing presents a conceptual framework for enterprise data exchange practices, strengthening trust in digital B2B relationships.

The digitalization of the energy sector introduces unique challenges. In Digital Transformation in Energy Sector: Cybersecurity Challenges and Implications, the risks associated with automation and the interconnectivity of critical infrastructure are explored, emphasizing the urgency of sector-specific solutions to ensure stability and security.

As fog computing becomes a key paradigm for real-time applications, Reliability and Security for Fog Computing Systems discusses strategies to ensure reliability and protection in these decentralized systems, often located at the network edge. Finally, the human factor remains a vital link in digital security. Improving the Cybersecurity Awareness of Young Adults through a Game-Based Informal Learning Strategy demonstrates that playful, informal learning approaches are effective in raising cybersecurity awareness among young people, an especially connected and vulnerable demographic.

This second edition reaffirms that digital privacy and security are not merely technical challenges

but also social, ethical, environmental, and educational ones. By bringing together contributions that span disciplines and contexts, this volume offers a comprehensive and updated view of the many fronts on which the fight for protection in the digital world is being waged. Building a secure digital future requires, more than ever, constant innovation, critical thinking, and multidisciplinary collaboration and cooperation.

José Braga de Vasconcelos, Hugo Barbosa, and Carla Cordeiro
Guest Editors

information

MDPI

Article

Optimized Digital Watermarking for Robust Information Security in Embedded Systems

Mohcin Mekhfioui [1,2,*], Nabil El Bazi [1], Oussama Laayati [1], Amal Satif [2], Marouan Bouchouirbat [1], Chaïmaâ Kissi [2], Tarik Boujiha [2] and Ahmed Chebak [1]

[1] Green Tech Institute (GTI), Mohammed VI Polytechnic University, Benguerir 43150, Morocco
[2] National School of Applied Sciences, Ibn Tofail University, Kenitra 14000, Morocco
* Correspondence: mohcin.mekhfioui@um6p.ma

Abstract: With the exponential growth in transactions and exchanges carried out via the Internet, the risks of the falsification and distortion of information are multiplying, encouraged by widespread access to the virtual world. In this context, digital image watermarking has emerged as an essential solution for protecting digital content by enhancing its durability and resistance to manipulation. However, no current digital watermarking technology offers complete protection against all forms of attack, with each method often limited to specific applications. This field has recently benefited from the integration of deep learning techniques, which have brought significant advances in information security. This article explores the implementation of digital watermarking in embedded systems, addressing the challenges posed by resource constraints such as memory, computing power, and energy consumption. We propose optimization techniques, including frequency domain methods and the use of lightweight deep learning models, to enhance the robustness and resilience of embedded systems. The experimental results validate the effectiveness of these approaches for enhanced image protection, opening new prospects for the development of information security technologies adapted to embedded environments.

Keywords: convolutional neural network; digital watermarking; image protection; cryptography; steganography; IoT; embedded systems

1. Introduction

Digital watermarking is a technique for embedding hidden information within digital media, like images, audio, or video, primarily used to protect intellectual property and verify content authenticity. Unlike visible watermarks in printed materials, digital watermarks are generally invisible to users and are integrated directly into the media data. This approach serves various purposes, such as copyright protection, content authentication, and tamper detection, while remaining robust against modifications, including compression and filtering [1]. Recent advancements have improved watermarking's resilience to digital processing, allowing embedded information to withstand common editing and transformation operations, thereby enhancing its effectiveness in securing digital assets [2].

By drawing on concepts from cryptography and signal processing, digital watermarking supports the tracking of intellectual property rights in a digitally pervasive environment, thus making it a crucial technology in today's multimedia and information-sharing age [3].

Optimizing digital watermarking for enhanced information security has become essential, especially in embedded environments where data integrity and protection against unauthorized access are critical. Contemporary research has developed sophisticated watermarking algorithms that balance robustness and invisibility by using advanced techniques

like the Shearlet domain and multi-objective optimization algorithms, improving resistance to hybrid attacks while preserving image quality [4].

To further enhance watermark security, algorithms incorporating artificial intelligence, such as artificial bee colony optimization, have been applied to strengthen watermark robustness and imperceptibility [5]. Other approaches leverage selective encryption combined with swarm optimization techniques to adjust embedding factors dynamically, thus improving the resilience of the watermark against various image processing attacks [6].

Emerging hybrid methods such as the integration of discrete wavelet transform (DWT) and singular value decomposition (SVD) are utilized to achieve high imperceptibility in medical image applications, critical for protecting sensitive healthcare data [7]. Furthermore, adaptive embedding techniques using neural networks and histogram analysis are being explored to optimize extraction processes, which is crucial in environments with the frequent transmission of data [8].

This evolving field also includes optimized solutions like the Lorenz chaotic function for adaptive watermark embedding in QR codes, enhancing security and preventing unauthorized data extraction [9]. For aerial remote sensing images, techniques combining redundant discrete wavelet transform and singular value decomposition have been implemented to provide robust protection in multimedia security [10].

Recent advancements in digital watermarking have greatly improved security and robustness, especially for applications requiring high imperceptibility and resilience against attacks. In the medical field, an optimized watermarking approach based on Integer Wavelet Transform and Particle Swarm Optimization has shown effectiveness in protecting sensitive images, balancing robustness and imperceptibility to suit various medical imaging modalities [11]. A novel approach integrating chaotic encryption for digital images has been proposed, enhancing the security and flexibility of watermark embedding through the Lorenz chaotic model [12]. Additionally, a hybrid technique combining dual watermarking and nature-inspired optimization algorithms, such as the Firefly and Particle Swarm Optimization methods, provides an optimal scaling factor for robust embedding in E-health applications [13]. Machine learning has also been applied to watermarking for improved performance; recent studies highlight its role in optimizing feature selection for secure, robust watermarking applications across various media [14]. In a parallel development, innovative watermarking schemes using Curvelet transform and multiple chaotic maps have demonstrated enhanced imperceptibility and localization capabilities, addressing both copyright protection and tampering detection effectively [15].

These advancements signify an important trend in watermarking, balancing imperceptibility, robustness, and security to protect data integrity in embedded environments, thereby reinforcing digital rights management and enhancing information security across various applications.

To give you a better idea of our contribution, Table 1 provides a comparative summary of the main existing digital watermarking techniques. It describes their main methods, strengths, and limitations, highlighting the diversity of approaches in this field and the need to find solutions optimized for embedded environments. This comparison highlights current research gaps, namely the absence of watermarking techniques that are both robust and lightweight for practical deployment in embedded systems with limited resources.

Table 1 highlights the diversity of digital watermarking approaches, ranging from classical SVD-based methods to hybrid techniques combining artificial intelligence, frequency domain transformations, and chaotic encryption. While many of these solutions demonstrate strong robustness or high imperceptibility, most suffer from limitations such as computational complexity, lack of adaptability, or unsuitability for embedded systems. These observations underline the need for an optimized method that balances security,

efficiency, and lightweight implementation—precisely the focus of the approach proposed in this article.

Table 1. Summary of major existing digital watermarking solutions.

Reference	Year	Techniques	Strengths	Limitations
Podilchuk & Delp [2]	2001	Algorithm taxonomy, applications	Foundational framework for watermarking systems	No empirical algorithm proposals
Chang et al. [1]	2005	SVD-based embedding	High imperceptibility and robustness to compression	Sensitive to geometric distortions
Sadiku et al. [3]	2017	Overview of watermarking types	General classification and use cases	No technical innovation or testing
Haghighi et al. [4]	2020	Shearlet, MLP, NSGA-II	Optimized blind and multipurpose watermarking	High computational complexity
Hemdan [7]	2020	SVD, DWT, wavelet fusion	High fidelity, secure for medical images	Increased processing due to scrambling
Sunesh et al. [8]	2020	ANN, histogram shape	Content-adaptive watermarking	Performance varies by image type
Sharma et al. [5]	2021	Nature-inspired optimization	Secure and robust for color images	Not ideal for grayscale images
Mohan et al. [6]	2021	Selective encryption, optimization	Robust transmission for natural images	Focused on landslide image applications
Pan et al. [9]	2021	Improved SMS, QR code embedding	Effective QR-specific watermarking	Not suitable for general images
Devi et al. [10]	2022	G-BAT hybrid optimization	Robust 3-level watermarking	Complex parameter tuning
Anand & Singh [13]	2022	Hybrid optimization and encryption	Tailored for E-healthcare	Highly domain-specific
Abdi & Boukli Hacene [11]	2023	Medical image optimization	Efficient and secure for E-health	Limited to medical data
Hao et al. [12]	2023	Chaotic encryption, live code	Real-time secure watermarking	Experimental; lacks benchmarks
Rai et al. [14]	2023	Machine learning	Adaptable, robust watermarking	Needs quality training data
Xiao et al. [15]	2023	Curvelet transform + multiple chaotic maps	High imperceptibility and precise localization	Complexity in implementation and parameter tuning

In this study, we developed and implemented a tattoo integrator on a Raspberry Pi platform, enabling the rapid incorporation of a user-selected tattoo. An optimization methodology was proposed, incorporating adapted arithmetic and a communication interface optimized for platform-specific hardware. A new software and hardware architecture dedicated to tattoo processing was introduced. The performance of the watermarking processor was analyzed, demonstrating that the designed system can run at high speed on Raspberry Pi. This adaptation illustrates the efficiency and flexibility of the Raspberry Pi platform for applications requiring rapid integration and high-performance execution.

2. Materials and Methods

2.1. System Architecture

This study proposes an innovative architecture for a watermarking system utilizing neural networks to ensure the security of multimedia data while maintaining robustness and energy efficiency. The developed system integrates several key components and processes (Figure 1):

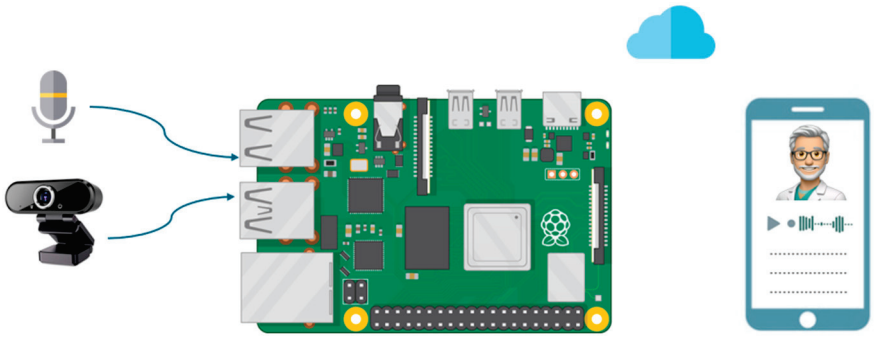

Figure 1. The architecture of the proposed watermarking system.

(a) Acquisition Block:

The acquisition block component captures the multimedia signals to be secured. It includes a camera for image acquisition and a microphone for audio capture. These signals are processed in real time by Raspberry Pi, which is equipped with interfaces to manage the input from these devices efficiently. The raw multimedia data are temporarily stored for further processing.

(b) Processing and Communication Block:

The processing and communication block is managed by Raspberry Pi, which integrates neural network algorithms to embed digital watermarks into the captured multimedia data.

- Watermark Embedding: Neural networks are trained to adaptively embed robust and imperceptible watermarks into images and audio signals while ensuring their fidelity.
- Data Encryption: To secure the watermarked multimedia, the data are encrypted before storage or transmission.
- Communication Module: The system establishes a secure connection to a web server, where the processed data are uploaded. This ensures that the data are accessible remotely for further decryption and validation [16].

(c) Decryption and Validation System:

The watermarked and encrypted multimedia data stored on the web server can be accessed and decrypted using an auxiliary embedded device or a smartphone application. This subsystem is responsible for validating the integrity of the watermark and ensuring data authenticity.

- Decryption Device: Either another embedded board (e.g., a secondary Raspberry Pi platform) or a smartphone processes the decryption and retrieves the embedded watermark.
- Verification Algorithms: Neural networks are also utilized at this stage to extract and validate the embedded watermark, ensuring robustness against potential attacks.

(d) Supervision and Monitoring System:

The Supervision System is an integral component of the proposed architecture, designed to provide users with a secure, efficient, and user-friendly platform for monitoring and managing watermarked multimedia data. This system is implemented as a cross-platform Android/iOS application developed using Python 3.13.0 frameworks, and it leverages AI algorithms to decrypt received images, extract embedded audio, and validate watermarks for data integrity. It enables the seamless playback of the recovered audio and includes a translation feature for converting the audio into the user's preferred language if needed. The application provides a secure environment with end-to-end encryption,

ensuring only authorized users can access or process the data, while real-time monitoring and alert features notify users of new files or anomalies.

This architecture emphasizes the seamless integration of watermarking techniques into a practical and secure multimedia management system, leveraging the computational power of neural networks and embedded hardware for robust, real-time operations.

2.2. Digital Watermarking Background

Image watermarking is a digital technique that involves embedding hidden information within an image to serve purposes such as copyright protection, authentication, and traceability. The primary objective is to embed a watermark that is imperceptible to viewers but resilient enough to survive various types of image manipulation, such as compression or cropping. This process involves balancing two key requirements: the watermark must be robust (resistant to removal or degradation from attacks) and visually imperceptible. By utilizing sophisticated embedding techniques, such as altering frequency or spatial domains, watermarking allows for durable and invisible security features within images [17].

Watermarking methods have advanced significantly to include adaptive schemes, which use the properties of the human visual system, such as masking effects, to further embed watermarks in less perceptible areas, making them robust to common attacks like JPEG compression and filtering [18]. Additionally, the spatial domain approach, which modifies the intensity of specific pixels, has also been effective in achieving robustness without requiring the original image for watermark extraction, ensuring security even under geometrical distortions [19]. Figure 2 summarizes a typical image watermarking flow, divided into three main steps: insertion, transmission, and extraction. In the insertion stage, a watermark (in our case an audio file) is inserted into a digital image using a secret authentication key. This step slightly modifies the original image, resulting in a watermarked version that looks almost identical to the original but contains hidden information.

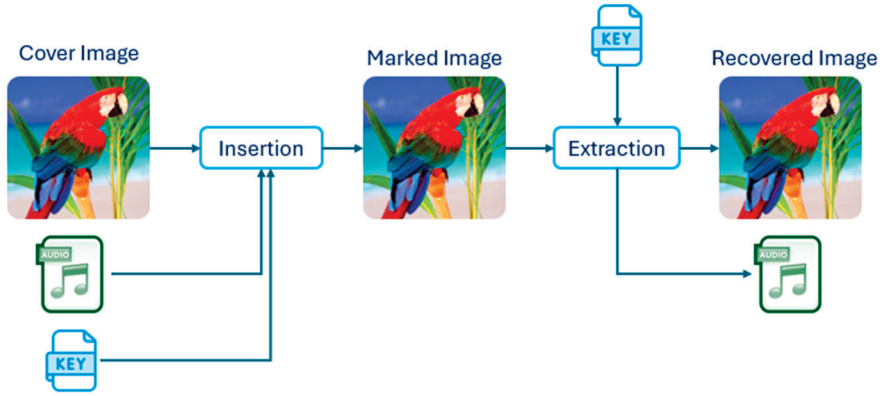

Figure 2. General schema of digital watermarking.

In the transmission phase, the watermarked image is transmitted over a wired or wireless transmission line. It may be exposed to various transformations such as compression, resizing, or noise.

In the extraction stage, the watermark is extracted from the image using the same authentication key as was used during insertion. The extracted watermark is then compared with the original. If it matches, the image is considered authentic. If not, the image may have been altered or modified.

2.2.1. Mathematical Principle of Watermarking

Watermarking methods can be categorized into two main types: spatial methods and frequency methods.

- Spatial Methods

Spatial methods embed the watermark directly in the time domain of the audio signal. A common approach is to add a low-amplitude signal to the original audio signal. Mathematically, this can be represented as follows:

$$y(t) = x(t) + \alpha w(t) \tag{1}$$

where $y(t)$ is the watermarked audio signal, $x(t)$ is the original audio signal, $w(t)$ is the watermark signal, and α is a weighting factor determining the amplitude of the watermark.

- Frequency Methods

Frequency methods, on the other hand, embed the watermark in the frequency domain of the audio signal. A common technique is to use the Fourier transform to convert the audio signal to the frequency domain, embed the watermark, and then convert the signal back to the time domain. This can be formulated as follows:

1. The Fourier transform of the original audio signal:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \tag{2}$$

2. Embedding the watermark in the frequency domain:

$$Y(f) = X(f) + \beta W(f) \tag{3}$$

3. Using the inverse Fourier transform to obtain the watermarked signal:

$$y(t) = F^{-1}\{Y(f)\} = \int_{-\infty}^{\infty} Y(f) e^{j2\pi ft} df \tag{4}$$

where ($X(f)$) and ($Y(f)$) are the frequency domain representations of the original and watermarked audio signals, respectively.

$W(f) = \int_{-\infty}^{\infty} w(t) e^{-j2\pi ft} dt$ is the watermark signal in the frequency domain.
β is a weighting factor.

2.2.2. Application

Watermarking technology has expanded into various applications, demonstrating its critical role in data protection, authentication, and privacy across diverse fields. In 5 G communication and the Internet of Things (IoT), watermarking secures data transmission by embedding identifiers that safeguard against unauthorized access—an essential feature given the extensive data exchange between interconnected devices [20]. Similarly, cyber systems and intellectual property protection leverage watermarking for digital rights management, ensuring traceability and the verification of digital content authenticity [21,22].

In medical imaging, watermarking serves as a privacy safeguard by embedding personal data within medical images without visually altering them, thus maintaining both privacy and data integrity during transmission and retrieval [23]. Moreover, smart city applications utilize watermarking to ensure data integrity in environments susceptible to cybersecurity threats, helping manage and secure large-scale information flow across urban infrastructures [24]. Watermarking's applications extend further into cloud storage, e-voting systems, and remote education, highlighting its integral role in maintaining security in increasingly interconnected digital ecosystems [25]. Figure 3 provides an overview of the key practical areas where digital watermarking is essential.

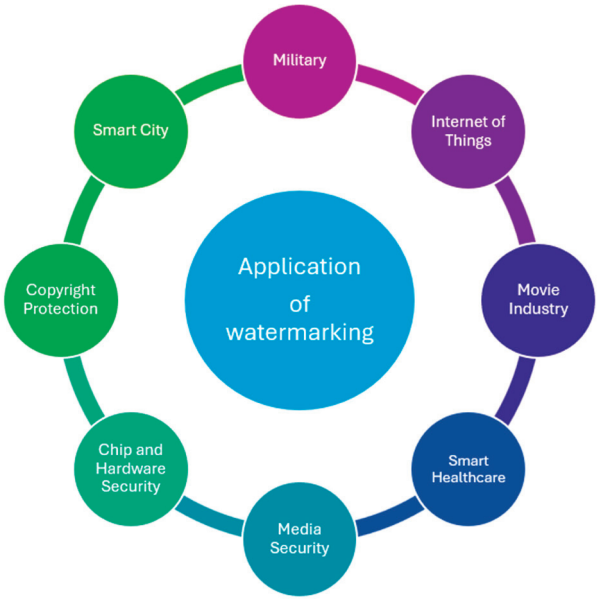

Figure 3. Applications for digital watermarking.

In the realm of audio data, watermarking presents unique challenges due to constraints involving sound quality and human perception. Research highlights that a watermark must be robust enough to endure compression and transmission operations without degrading audio quality. Optimization techniques are essential to balance robustness with imperceptibility, ensuring that the watermark remains unobtrusive while maintaining its functionality.

2.2.3. Performance Requirements in Digital Watermarking

For an effective watermarking scheme, the design algorithm must fulfill specific requirements and characteristics, as these are essential for assessing the technology's performance. The importance of each characteristic varies based on the intended application of the watermark. Below, we outline the key requirements for a digital watermarking scheme along with the evaluation metrics used for each.

(a) Robustness

Robustness measures the watermark's ability to withstand various signal-processing manipulations, including both intentional attacks (e.g., compression, noise) and unintentional distortions. Different types of watermarking schemes—such as robust, fragile, and semi-fragile—target varying levels of robustness. In instances where the watermark takes the form of a two-dimensional matrix, its resilience is often assessed via the normalized cross-correlation (NC), measuring the similarity between the original watermark w and the extracted watermark w' [26]:

$$NC = \frac{\sum_{i=1}^{H}\sum_{j=1}^{L}\left(w_{ij}w'_{ij}\right)}{\sqrt{\sum_{i=1}^{H}\sum_{j=1}^{L}\left(w_{ij}\right)^2}\sqrt{\sum_{i=1}^{H}\sum_{j=1}^{L}\left(w'_{ij}\right)^2}} \quad (5)$$

where H and L define the height and width of the watermark.

The image watermarking techniques that are robust against various geometric and filtering attacks have shown improved robustness through adaptive embedding methods [27].

(b) Imperceptibility

Imperceptibility ensures that the watermark remains visually or audibly undetectable, preserving the quality of the host content. This is particularly essential in images, audio, and video media.

- The *Peak Signal-to-Noise Ratio (PSNR)* and *Structural Similarity Index Model (SSIM)* are common metrics; *PSNR* values above 40 dB are generally considered imperceptible [28].

$$PSNR = 10 \times log_{10}\left(\frac{\max(c)^2}{\frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C}\left(c_{ij}-m_{ij}\right)^2}\right) \quad (6)$$

  where *max(c)* is the largest possible pixel value for the cover image c, which is 255 if we use 8 bits for each grayscale value, and R and C denote the height and width of images c and m.

- Adaptive techniques, which adjust watermark embedding based on image features, have shown improved imperceptibility, as they use human visual sensitivity to reduce visible distortion [29].

(c) Capacity

Capacity represents the maximum amount of data that can be embedded as a watermark without compromising quality. Applications requiring extensive metadata or copyright information embedded within content prioritize high-capacity methods.

- Higher capacities often necessitate trade-offs with imperceptibility and robustness. Multiple watermark systems are increasingly employed to balance these demands [30].

(d) Complexity

Complexity addresses the computational efficiency needed for watermark embedding and extraction, especially vital for real-time applications like video streaming.

- Efficient processing times are crucial, often requiring optimization through hardware implementations or advanced algorithms [31].

(e) Security

Security ensures that the watermark resists removal or alteration, making encryption and secure embedding techniques essential.

- Measures like the *NPCR (Number of Changing Pixel Rate)* and *UACI (Unified Averaged Changed Intensity)* are used to assess resilience against tampering [32]. These two measures are used to evaluate the efficiency of image watermarking against potential attacks. They are usually used to analyze the resistance of the watermarked images to pixel-level changes. *NPCR* and *UACI* scores should always be close to 1 and 0.33, respectively, to achieve good security. Higher values mean resistance will be better.

2.2.4. Watermarking with Neural Network

Watermarking with neural networks has emerged as a robust method for protecting intellectual property in multimedia and AI models. Various techniques have been developed, such as embedding watermarks directly into the outputs of neural networks by training them alongside a watermark extraction network, which ensures that the watermark is integrated without impairing the network's task performance [33]. Other approaches, like dynamic watermarking, leverage neural networks to adaptively embed and retrieve watermarks, increasing resilience against attacks such as false authentication and image compression [34]. Additionally, digital watermarking techniques embed watermarks into deep neural network parameters, ensuring robustness against fine-tuning and parameter pruning while retaining model performance [35]. Blind watermarking algorithms also

demonstrate efficacy by using back-propagation neural networks to embed watermarks that are resilient to various image processing attacks [36]. These neural network-based watermarking techniques highlight their adaptability and robustness, making them essential tools for protecting digital assets and verifying model ownership in the era of AI.

2.2.5. Deep Learning-Based Watermark Embedding and Extraction Architecture

To improve robustness while maintaining a lightweight computational profile suitable for embedded systems, we designed a deep learning-based watermarking system using convolutional neural networks (CNNs). This approach allows for the embedding of an audio signal into an image with high imperceptibility and efficient extraction capabilities.

The proposed system includes two core components: a hiding model and an extraction model. The hiding model takes two inputs: a color image (of size $128 \times 128 \times 3$) and an audio signal reshaped into a grayscale matrix of the same spatial dimensions ($128 \times 128 \times 1$). These two inputs are concatenated along the channel axis to form a unified input volume, which then passes through two convolutional layers, each with 64 filters and a 3×3 kernel using ReLU activation. The final output layer uses a 3-filter convolution with a sigmoid activation function to generate a stego-image, which visually resembles the original image but carries the embedded audio data in its pixel values.

The extraction model, in turn, is designed to reverse this process. It takes the stego-image as input and passes it through a similar set of convolutional layers to decode and reconstruct the embedded audio matrix. The output layer has a single filter and uses a sigmoid activation function to produce the final extracted audio signal.

Training was conducted using a small dataset for proof of concept, consisting of a single image–audio pair. The model was trained for 100 epochs with a batch size of 1 using the Adam optimizer and a learning rate of 0.0001. Although our neural network learns to embed and extract data directly from pixel patterns, this approach is conceptually compatible with LSB-style embedding, as it alters pixel values subtly in a way that mimics the minimal perturbation found in LSB techniques while leveraging learned patterns for greater robustness. Despite the minimal dataset, the results demonstrated the successful embedding and extraction of the audio signal with good visual imperceptibility and waveform fidelity.

3. Results

This section presents the results obtained when evaluating the watermarking system developed to embed a .wav audio file in a PNG image using Raspberry Pi. Tests were carried out to evaluate performance in terms of processing time, data quality after watermarking, and robustness in the face of various modifications.

3.1. Simulation Results

To evaluate the performance of the proposed AI-based watermarking process, we used a 1920×1080-pixel image and a 5 s audio extract. The image was chosen from a collection of standard images, while the audio was recorded in WAV format, with a sampling frequency of 44.1 kHz and a bit depth of 16 bits. These data were chosen for their ability to represent common multimedia formats while also enabling us to evaluate the effectiveness of the watermarking system on relatively large files.

3.1.1. Processing Time

The system's overall processing time was measured from the data capture, processing, and transmission stages. Capturing audio data using a microphone connected to Raspberry Pi takes around 0.7 s for a 5 s .wav file. Similarly, capturing a PNG image with a camera and resizing it to a standard resolution of 1920×1080 pixels takes an average of 0.2 s.

Integrating the audio file into the image using the least significant bit (LSB) watermarking algorithm takes around 1.2 s for typical audio file sizes. Finally, the transmission of the marked image to the web server via a local Wi-Fi network takes an average of 0.8 s. This means that the entire process, from data capture to transmission, takes less than 2.9 s, demonstrating the viability of the method for real-time applications.

Figure 4 summarizes the processing times associated with each step. Among all the stages, the watermark embedding process is the most time-consuming, which is consistent with the computational effort required to modify pixel values at scale, especially for high-resolution images. Although the LSB technique is conceptually simple, its implementation over a 1920 × 1080-pixel matrix requires sequential operations that accumulate to over one second of processing.

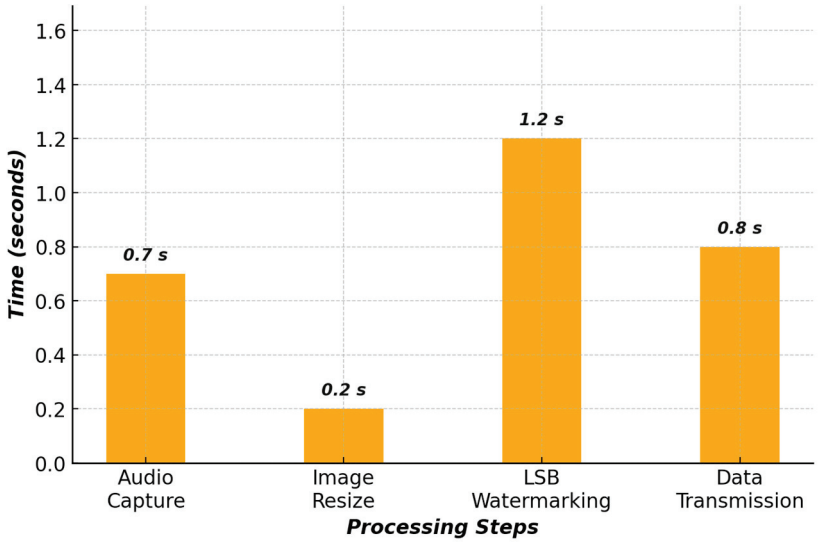

Figure 4. Processing times for different steps.

In contrast, image acquisition and resizing are fast operations due to the small file size and the optimized pipeline between the CSI camera and the processor. Audio capture is slightly more time-intensive, as the system buffers and encodes several seconds of audio in WAV format. Transmission time depends on the size of the output image and the stability of the Wi-Fi network but remains under one second in all tested scenarios.

3.1.2. Data Quality

Image and audio quality after watermarking was assessed to ensure the efficiency of the process. The quality of the marked image was analyzed by measuring the Peak Signal-to-Noise Ratio (PSNR), which reached an average value of 35.86 dB. This value indicates that the modifications made to the image are imperceptible to the human eye, thus guaranteeing the discretion of the watermarking.

For extracted audio, quality was measured using the Signal-to-Noise Ratio (SNR), which reached an average of 37.2 dB. These results show that the extracted audio is clear and intelligible, with negligible losses compared to the original file. These two measurements confirm that the system can preserve data quality while performing watermarking.

Figure 5 shows the PSNR values for five different test images after watermarking. All values lie between 34.4 dB and 37.3 dB, confirming that image fidelity is consistently preserved across different image types. Slight variations in the PSNR are mainly due to differences in image texture and color complexity; for example, images with flat or uniform areas (such as a wall or a white table) tend to retain slightly higher PSNR values, while

those with more detailed textures (such as nature scenes) show a slightly lower PSNR due to greater embedding activity at the pixel level.

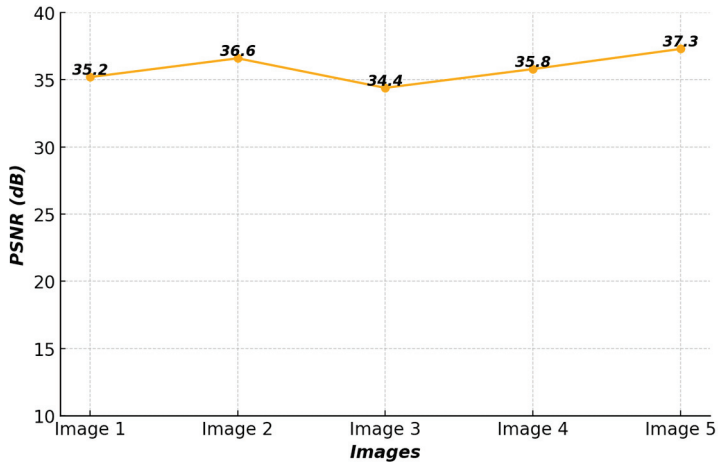

Figure 5. PSNR values for images after watermarking.

Nevertheless, all measured values remain above the 30 dB threshold, which is widely accepted as the point beyond which distortions are considered imperceptible to the human eye. This proves that the system achieves its objective of invisible watermarking, ensuring that the visual quality of content is not affected for practical use.

3.1.3. Robustness

The system was tested to assess its robustness in the face of various transformations applied to the marked image, such as compression, geometric transformations, and the addition of noise.

- The system is robust to lossless compression (PNG), with minimal degradation in the quality of the extracted audio data.
- Lossy compressions, such as those associated with the JPEG format, significantly reduce the quality of extracted data, especially at high compression ratios.
- Minor geometric transformations (e.g., rotation of less than 5° or partial cropping) do not significantly affect performance. However, major transformations, such as a 90° rotation or excessive cropping, make correct data extraction difficult.
- Finally, the addition of moderate noise in the image slightly affects the extracted audio, but it remains intelligible.

These results are visually summarized in Figure 6, which shows the Signal-to-Noise Ratio (SNR) of the extracted audio under each condition. The SNR remained above 30 dB in cases of minor distortions, confirming robustness and clear audio output. In contrast, JPEG compression and strong rotations caused a drop below 30 dB, showing the watermark's vulnerability under these more aggressive modifications.

In summary, the watermarking system we developed offers satisfactory performance for practical applications. Its fast processing time (less than 2.9 s) makes it suitable for embedded environments and scenarios requiring real-time processing. The high quality of the images and audio after watermarking guarantees the system's discretion and efficiency. Finally, although the system is robust to minor modifications and lossless compression, there is room for improvement to increase its resistance to lossy compression and complex geometric transformations. These results demonstrate that the system is a viable solution for applications requiring secure embedded data integration.

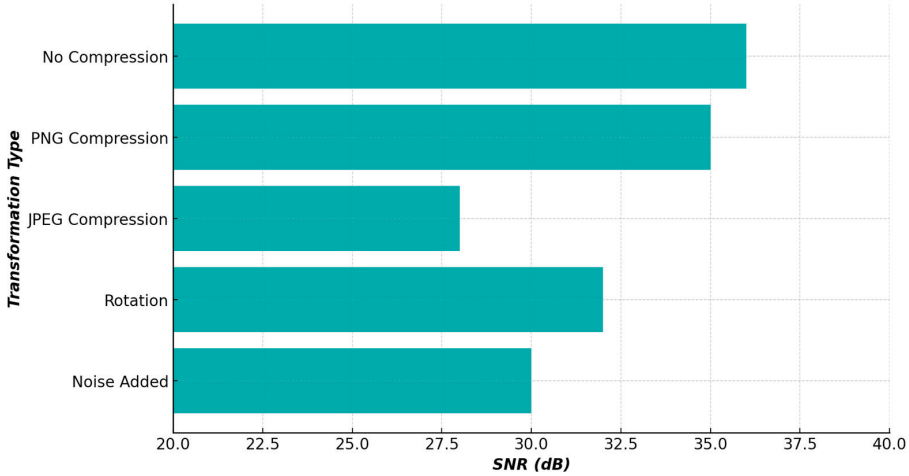

Figure 6. The robustness of the watermarking system (SNR).

3.2. Real-Time Results

The proposed system was designed, assembled, and tested in a laboratory environment. This section describes the hardware configuration and connections between the various components.

The assembly was designed to provide a stable and functional configuration. A camera was connected to the Raspberry Pi platform's CSI port, while a USB microphone [36] was plugged into one of the available USB ports. All components were installed on a prototyping board, with support to stabilize the camera. Raspberry Pi was powered by a 5 V/3A AC adapter to ensure reliable operation. A local Wi-Fi network connection was configured to enable the transmission of marked data to the server.

Figure 7 shows a block diagram of the real system, showing the connections between the various components.

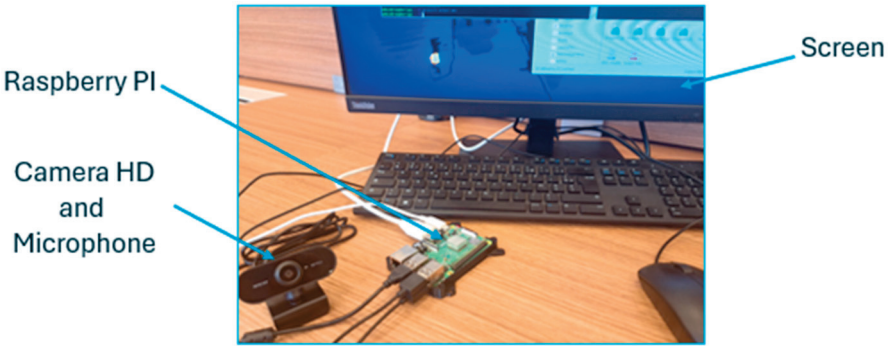

Figure 7. The hardware configuration of our system.

The process was divided into several stages. Firstly, the microphone and camera were tested to capture real-time data, namely audio files in .wav format and PNG images. Next, the audio was embedded in the image locally on Raspberry Pi using the least significant bit (LSB)-based watermarking algorithm. The figure below illustrates how the least significant bits (LSBs) are modified in the watermarking process. The image is compressed to reduce file size, and metadata such as the image format or unique identifier are included. Transmission is conducted via HTTP protocol, using a Python library as requested. To guarantee data security during transfer, an HTTPS connection is preferred. The web server, configured to receive these files, saves them and prepares them for further processing. This architecture enables efficient, secure communication between Raspberry Pi and the server.

This system offers several advantages. The components used are cost-effective, making the system accessible to a wide range of environments. The configuration is flexible, allowing sensors to be added or replaced for other applications. Finally, the system's simplicity makes it easy to transport, making it suitable for scenarios requiring rapid deployment.

Data Decoding and Validation

Once the marked image has been received by the server, a dedicated script extracts the embedded audio data. This process inverts the LSB algorithm, reading the modified bits from the image pixels to reconstruct the binary sequence of the audio file. The audio is then recreated and saved in its original .wav format.

To validate data integrity, a comparison is made between the extracted audio and the original audio. This allows us to assess the fidelity of the integration and extraction process. Image quality is also analyzed to ensure that watermarking is imperceptible and does not degrade the visual content.

We also developed a mobile application compatible with iOS and Android to retrieve the data generated by our system. This application aims to make the data more accessible and easier to manage and interpret while enabling real-time monitoring [37]. It allows users to view the watermarking results by extracting the audio message and the original image. Additionally, it offers the ability to translate the message into multiple languages. This application was developed in Python and consists of three main pages.

Authentication Page: The authentication page allows the user to log in to the application to access its features by entering their username and password.

Real-Time Monitoring Page: The real-time monitoring page displays a variety of information related to the last marked image received, followed by the extracted audio message, its transcription, and its translation into English.

History Page: The history page shows the history of the marked images received, as well as the extracted audio and the recovered image.

Figure 8 illustrates real-time result tracking on the mobile interface. Once the marked image reaches the web server, it is integrated into the application, and an embedded script extracts the audio file and the original image using a decryption key stored in the application's database. Then, another script handles the transcription of the audio file and displays the message in the designated area before translating it into English or another language predefined during account creation. Once the process is complete, all data are saved in history for future reference.

To secure access to embedded audio data, the system uses symmetrical AES-128 encryption to encode the audio content before inserting it into the image. This encryption method was chosen in view of AES's robust security features, high-speed performance, and minimal processing requirements, making it well suited for real-time use on embedded equipment such as Raspberry Pi.

A private key for decryption is stored in a database internal to the mobile application and is only accessible to logged-in users of the mobile application. After receiving a watermarked image from the web server, the mobile application automatically retrieves the decryption key, which is used to extract the hidden audio content and decode it using an integrated script.

In summary, this methodology offers a consistent and robust workflow for embedding an audio file in an image, using Raspberry Pi as the embedded processing platform. The system exploits simple hardware and software tools while guaranteeing secure transmission and accurate data processing on a remote server.

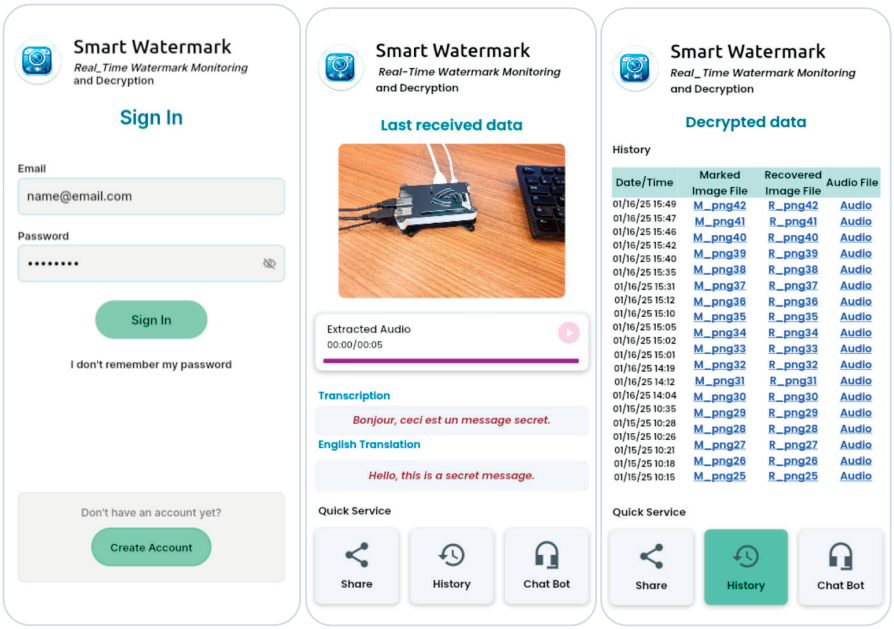

Figure 8. The watermarking results on the mobile interface.

Although the proposed watermarking system demonstrates high performance and practical viability for embedding audio into images in embedded environments, it also introduces critical privacy and legal implications, especially when the audio content includes personal or sensitive information. In scenarios where the embedded audio may contain identifiable voice data or confidential messages, the system must adhere to data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union. These frameworks emphasize key principles like data minimization, transparency, and the necessity of explicit user consent before processing personal data [38].

The unauthorized embedding of personal audio could potentially lead to privacy violations or legal disputes, particularly in applications such as E-health, smart surveillance, or IoT-based communication systems, where sensitive or private information is involved. To address these concerns, future implementations should integrate user consent mechanisms, data encryption, and possibly anonymization strategies to protect identities and comply with applicable laws. Furthermore, the ethical deployment of such AI-driven systems must account not only for technical robustness but also for broader social implications. Promoting transparency and user control over data embedding and extraction processes is key to ensuring responsible innovation in secure multimedia watermarking [39].

4. Conclusions

This paper presents an AI-based watermarking system capable of embedding and extracting audio signals within images, implemented and tested on a Raspberry Pi platform. The system demonstrated efficient processing times, with an average execution time under 2.9 s, validating its feasibility for real-time applications in embedded environments.

Quantitative evaluation showed high image fidelity, with Peak Signal-to-Noise Ratios (PSNRs) remaining above acceptable thresholds, indicating minimal perceptual distortion. The extracted audio also maintained a clear signal, confirming good transmission integrity under minor distortions and lossless compression.

Real-time tests confirmed the system's ability to integrate seamlessly with mobile applications, offering an intuitive user interface for retrieving and decoding audio-embedded messages. Thanks to its low-cost hardware and modular design, the system shows strong

potential for use in applications such as secure data transmission, multimedia authentication, and IoT-based messaging.

Future work will focus on enhancing resistance to lossy compression and complex geometric attacks, as well as integrating stronger encryption mechanisms to improve data confidentiality. These improvements aim to make the system even more robust for secure and scalable deployment in real-time multimedia watermarking scenarios.

Author Contributions: Conceptualization, M.M.; Methodology, M.M. and N.E.B.; Software, M.M. and C.K.; Validation, T.B. and A.C.; Formal analysis, M.M., N.E.B., O.L., A.S., M.B., C.K., T.B. and A.C.; Investigation, M.M., N.E.B. and O.L.; Resources, M.M. and A.C.; Data curation, M.M.; Writing—original draft, M.M.; Writing—review & editing, M.M. and N.E.B.; Visualization, O.L., A.S., M.B., C.K. and T.B.; Supervision, C.K., T.B. and A.C.; Project administration, A.C.; Funding acquisition, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NC	Normalized cross-correlation
CNN	Convolutional neural network
SVD	Singular value decomposition
DWT	Discrete wavelet transform
SSIM	Structural Similarity Index Model
IoT	Internet of Things
PSNR	Peak Signal-to-Noise Ratio
SNR	Signal-to-Noise Ratio
LSB	Least significant bit
NPCR	Number of Changing Pixel Rate
UACI	Unified Averaged Changed Intensity

References

1. Chang, C.; Tsai, P.; Lin, C.-C. SVD-based digital image watermarking scheme. *Pattern Recognit. Lett.* **2005**, *26*, 1577–1586. [CrossRef]
2. Podilchuk, C.; Delp, E. Digital watermarking: Algorithms and applications. *IEEE Signal Process. Mag.* **2001**, *18*, 33–46. [CrossRef]
3. Sadiku, M.; Shadare, A.E.; Musa, S. Digital watermarking. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2017**, *7*, 414. [CrossRef] [PubMed]
4. Haghighi, B.B.; Taherinia, A.; Harati, A.; Rouhani, M. WSMN: An optimized multipurpose blind watermarking in Shearlet domain using MLP and NSGA-II. *Appl. Soft Comput.* **2020**, *101*, 107029. [CrossRef]
5. Sharma, S.; Sharma, H.; Sharma, J.B.; Poonia, R. A secure and robust color image watermarking using nature-inspired intelligence. *Neural Comput. Appl.* **2021**, *35*, 4919–4937. [CrossRef]
6. Mohan, A.; Anand, A.; Singh, A.; Dwivedi, R.; Kumar, B. Selective encryption and optimization based watermarking for robust transmission of landslide images. *Comput. Electr. Eng.* **2021**, *95*, 107385. [CrossRef]
7. Hemdan, E.E. An efficient and robust watermarking approach based on single value decompression, multi-level DWT, and wavelet fusion with scrambled medical images. *Multimed. Tools Appl.* **2020**, *80*, 1749–1777. [CrossRef]
8. Sunesh, R.; Kishore, R.; Saini, A. Optimized image watermarking with artificial neural networks and histogram shape. *J. Inf. Optim. Sci.* **2020**, *41*, 1597–1613. [CrossRef]

9. Pan, J.-S.; Sun, X.-X.; Chu, S.; Abraham, A.; Yan, B. Digital watermarking with improved SMS applied for QR code. *Eng. Appl. Artif. Intell.* **2021**, *97*, 104049. [CrossRef]
10. Devi, K.J.; Singh, P.; Dash, J.; Thakkar, H.; Santamaría, J.; Krishna, M.V.J.; Romero-Manchado, A. A new robust and secure 3-level digital image watermarking method based on G-BAT hybrid optimization. *Mathematics* **2022**, *10*, 3015. [CrossRef]
11. Abdi, H.; Boukli Hacene, I. An optimized medical image watermarking approach for E-health applications. *Med. Technol. J.* **2023**, *5*, 594–603. [CrossRef]
12. Hao, W.; Wei, X.; Zhang, W.; Xie, R. Live code digital watermarking technology based on chaotic encryption. In Proceedings of the 2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 16–18 June 2023. [CrossRef]
13. Anand, A.; Singh, A.K. Hybrid nature-inspired optimization and encryption-based watermarking for E-healthcare. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 2033–2040. [CrossRef]
14. Rai, M.; Hemlata. Robust digital watermarking based on machine learning. In Proceedings of the 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 18–20 October 2023. [CrossRef]
15. Xiao, Y.; Xu, Y.-C.; Zhou, N.-R.; Lin, Z.-R. Digital watermarking scheme based on curvelet transform and multiple chaotic maps. *Opt. Appl.* **2023**, *53*, 291–305. [CrossRef]
16. Mekhfioui, M.; Benahmed, A.; Chebak, A.; Elgouri, R.; Hlou, L. The Development and Implementation of Innovative Blind Source Separation Techniques for Real-Time Extraction and Analysis of Fetal and Maternal Electrocardiogram Signals. *Bioengineering* **2024**, *11*, 512. [CrossRef] [PubMed]
17. Hsu, C.T.; Wu, J.L. Hidden digital watermarks in images. *IEEE Trans. Image Process.* **1999**, *8*, 58–68. [CrossRef]
18. Swanson, M.D.; Zhu, B.; Tewfik, A.H. Transparent robust image watermarking. In Proceedings of the International Conference on Image Processing, Lausanne, Switzerland, 16–19 September 1996; Volume 3, pp. 211–214. [CrossRef]
19. Nikolaidis, A.; Pitas, I. Robust image watermarking in the spatial domain. *Signal Process.* **1998**, *66*, 385–403. [CrossRef]
20. Li, H.; Zhang, W.; Sun, Y. IoT and 5G communication watermarking techniques. *Commun. Digit. Secur.* **2019**, *7*, 32–47.
21. Smith, T.; Williams, M.; Lee, D. Digital rights management in cyber systems via watermarking. *Cybersecur. Innov. J.* **2020**, *15*, 211–225.
22. Lansari, M.; Bellafqira, R.; Kapusta, K.; Thouvenot, V.; Bettan, O.; Coatrieux, G. When Federated Learning Meets Watermarking: A Comprehensive Overview of Techniques for Intellectual Property Protection. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1382–1406. [CrossRef]
23. Jones, P.; Wang, Q. Protecting patient privacy in medical imaging through watermarking. *Healthc. Data J.* **2021**, *5*, 99–112.
24. Brown, A.; Chen, X.; Zhao, L. Smart city data integrity and security with watermarking. *J. Urban Comput.* **2018**, *12*, 145–158.
25. Chen, L.; Zhao, Y. Watermarking for secure cloud storage and e-governance applications. *Int. J. Cloud Secur.* **2019**, *8*, 78–89.
26. Chen, B.; Wornell, G. Achievable performance of digital watermarking systems. In Proceedings of the IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, 7–11 June 1999; Volume 1, pp. 13–18. [CrossRef]
27. Qi, X.; Qi, J. A robust content-based digital image watermarking scheme. *Signal Process.* **2007**, *87*, 1264–1280. [CrossRef]
28. Akter, A.; Ullah, M. Digital Watermarking with a New Algorithm. *Int. J. Res. Eng. Technol.* **2014**, *3*, 212–217. [CrossRef]
29. Qi, H.; Zheng, D.; Zhao, J. Human visual system based adaptive digital image watermarking. *Signal Process.* **2008**, *88*, 174–188. [CrossRef]
30. Zhang, F.; Zhang, X. Performance Evaluation of Multiple Watermarks System. In Proceedings of the Second Workshop on Digital Media and Its Application in Museum & Heritages (DMAMH 2007), Chongqing, China, 10–12 December 2007; pp. 15–18. [CrossRef]
31. Roy, S.; Li, X.; Shoshan, Y.; Fish, A.; Yadid-Pecht, O. Hardware Implementation of a Digital Watermarking System for Video Authentication. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 289–301. [CrossRef]
32. Garg, P.; Kishore, R. Performance comparison of various watermarking techniques. *Multimed. Tools Appl.* **2020**, *79*, 25921–25967. [CrossRef]
33. Wu, H.; Liu, G.; Yao, Y.; Zhang, X. Watermarking Neural Networks With Watermarked Images. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2591–2601. [CrossRef]
34. Gu, T.; Li, X. Dynamic digital watermark technique based on neural network. In *Independent Component Analyses, Wavelets, Unsupervised Nano-Biomimetic Sensors, and Neural Networks VI*; SPIE: Bellingham, WA, USA, 2008.
35. Uchida, Y.; Nagai, Y.; Sakazawa, S.; Satoh, S. Embedding Watermarks into Deep Neural Networks. In Proceedings of the ACM International Conference on Multimedia Retrieval, Bucharest, Romania, 6–9 June 2017.
36. Huang, S.; Zhang, W.; Feng, W.; Yang, H. Blind watermarking scheme based on neural network. In Proceedings of the World Congress on Intelligent Control and Automation, Chongqing, China, 25–27 June 2008.
37. Mekhfioui, M.; Elgouri, R.; Satif, A.; Hlou, L. Real-time implementation of a new efficient algorithm for source separation using matlab & arduino due. *Int. J. Sci. Technol. Res.* **2020**, *9*, 4.

38. Voigt, P.; von dem Bussche, A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*; Springer: Berlin/Heidelberg, Germany, 2017. [CrossRef]
39. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]

information

MDPI

Article

A Comparative Study of Privacy-Preserving Techniques in Federated Learning: A Performance and Security Analysis

Eman Shalabi [1], Walid Khedr [1,2], Ehab Rushdy [1] and Ahmad Salah [3,*]

[1] Faculty of Computers & Informatics, Zagazig University, Zagazig 44519, Egypt; emanselem@zu.edu.eg (E.S.); wkhedr@taibahu.edu.sa (W.K.); ehab.rushdy@zu.edu.eg (E.R.)
[2] College of Computer Science and Engineering, Taibah University, Yanbu 966144, Saudi Arabia
[3] College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri P.O. Box 466, Ad Dhahirah, Oman
* Correspondence: ahmad.salah@utas.edu.om

Abstract: Federated learning (FL) is a machine learning technique where clients exchange only local model updates with a central server that combines them to create a global model after local training. While FL offers privacy benefits through local training, privacy-preserving strategies are needed since model updates can leak training data information due to various attacks. To enhance privacy and attack robustness, techniques like homomorphic encryption (HE), Secure Multi-Party Computation (SMPC), and the Private Aggregation of Teacher Ensembles (PATE) can be combined with FL. Currently, no study has combined more than two privacy-preserving techniques with FL or comparatively analyzed their combinations. We conducted a comparative study of privacy-preserving techniques in FL, analyzing performance and security. We implemented FL using an artificial neural network (ANN) with a Malware Dataset from Kaggle for malware detection. To enhance privacy, we proposed models combining FL with the PATE, SMPC, and HE. All models were evaluated against poisoning attacks (targeted and untargeted), a backdoor attack, a model inversion attack, and a man in the middle attack. The combined models maintained performance while improving attack robustness. FL_SMPC, FL_CKKS, and FL_CKKS_SMPC improved both their performance and attack resistance. All the combined models outperformed the base FL model against the evaluated attacks. FL_PATE_CKKS_SMPC achieved the lowest backdoor attack success rate (0.0920). FL_CKKS_SMPC best resisted untargeted poisoning attacks (0.0010 success rate). FL_CKKS and FL_CKKS_SMPC best defended against targeted poisoning attacks (0.0020 success rate). FL_PATE_SMPC best resisted model inversion attacks (19.267 MSE). FL_PATE_CKKS_SMPC best defended against man in the middle attacks with the lowest degradation in accuracy (1.68%), precision (1.94%), recall (1.68%), and the F1-score (1.64%).

Keywords: federated learning; privacy preserving; homomorphic encryption; secure multi-party computation; private aggregation of teacher ensembles; poisoning attacks; backdoor attack; model inversion attack; man in the middle attack

1. Introduction

In an increasingly interconnected world, federated learning (FL) is a novel machine learning (ML) approach that addresses the issues of decentralization, security, and data privacy. FL is an essential approach for distributed ML, especially because it protects user privacy and improves communication effectiveness. Because of this, it is extremely essential in the data-driven world of today, where privacy issues are critical [1]. FL is

extensively utilized in several domains such as the Internet of Things (IoT), computer vision (CV), natural language processing (NLP), healthcare, financial services, and medical image analyses [2,3].

This approach allows models to be trained on distributed datasets without the need to centralize the data. It permits several parties to cooperatively train a common model without sharing their raw data, in contrast to conventional ML techniques [4]. In FL, user data are not shared with a central server; only the parameters of ML models are [5]. It can achieve great learning performance for the benefit of the client while simultaneously protecting sensitive client data and also gives customers the ability to retain control over their personal data in an environment where data breaches and privacy violations are common [6].

Despite all the advantages of FL in enhancing data privacy, FL is still vulnerable to privacy risks: the local models' updates may disclose sensitive information about participants, attackers may still be able to obtain private data even in cases when only gradient data are transferred to central servers, attackers may infer and ultimately reconstruct private data from model updates shared between participants and the central server, an adversary or a malicious client may poison the training process by modifying local training data or gradients, affecting the shared model's integrity, adversaries may access sensitive information by intercepting transferred data during model updates, and so on [3,7].

According to several studies, including [2,8,9], to handle the above problems of privacy concerns in FL, several privacy-preserving approaches have been proposed, such as differential privacy (DP) [10], homomorphic encryption (HE) [11], and Secure Multi-Party Computation (SMPC) [12]. The Private Aggregation of Teacher Ensembles (PATE) is a DP technique. The PATE involves a combination of multiple teacher models and one student model [13]. Each teacher model is trained on a separate subset of the data [14].

Using a Laplace or Gaussian mechanism, noise is added to the aggregated results of teacher models [15]. The student model is trained on the aggregated labels from teacher models and then used for predictions [16]. SMPC enables the cooperative computation of a function over private inputs by several parties without disclosing those data, so it ensures confidentiality by safely combining model updates in FL without disclosing individual updates to the central server [17,18].

HE is a cryptographic technique that permits calculations on an encrypted form of data called ciphertext, producing an encrypted result that, upon decryption, matches the outcome of operations conducted on the original form of the data, called plaintext [19,20]. In FL, each client only encrypts their local model using HE so that the aggregator can perform mathematical operations on it [21,22]. There have been studies that integrated one privacy-preserving technique with FL and other studies that combined more than one privacy-preserving technique with FL.

The studies in [13,23,24] integrated the PATE with FL. In [24], the authors proposed the FREDY framework, which combines the PATE with FL and applies Convolutional Neural Networks. Through the federated training of multiple teacher models, inference on publicly accessible unlabeled data, prediction aggregation with Laplace noise, and the training of a student model on the labeled data resulting from this procedure, the PATE was incorporated into FREDY. The CIFAR10 and MNIST datasets were used. The work in [25–27] combined SMPC with FL.

The authors of [25] proposed the PrivatEyes framework, which is an FL-based framework for the protection of gaze estimation data. PrivatEyes integrates SMPC with FL and applies Convolutional Neural Networks. The MPIIFaceGaze, GazeCapture, and NVGaze datasets were used for evaluation. Several research studies such as [11,28,29] integrated HE with FL. The authors of [28] combined HE with FL for intrusion detection in the Internet of

Vehicles with limited computing resources. HE is used by Vehicle Users for the encryption of offloaded data for the purpose of transmitting it to the server. Vehicle Users also use HE to encrypt their local model updates before sending them to the central server to perform an aggregation process. A Deep Neural Network and the Edge-IIoT dataset are used.

The authors of [20,30,31] proposed combining FL with HE and SMPC. To handle users' dropout, particularly in environments with limited resources, [30] integrated FL, HE, and SMPC. Multi-homomorphic encryption was used to perform computations on encrypted data without decrypting it. Secret sharing was used for SMPC. To handle user dropouts, secret sharing was used, along with a random mask secure aggregation mechanism that preserved the masks' privacy. The aggregation process only needed two communication rounds, and its communication and computational overheads increased linearly with the number of users.

The work in [32] combined FL, HE, and DP. In [32], the authors integrated FL, HE, and DP to propose the FLCP framework. FLCP uses a Convolutional Neural Network. Following the computation of the local model updates, clients perform weight compression using the AWC-FedAvg algorithm. DP is applied through the addition of noise to the compressed local model updates using a Laplace mechanism. Using HE, the noisy compressed updates are encrypted. The clients transmit encrypted updates to the central server for the aggregation and updating of the global model. The MNIST and CIFAR-10 datasets are used for evaluation.

Despite the growing interest in FL security mechanisms, the literature is notably deficient in in-depth comparative studies of different combinations with different security levels. While individual privacy-preserving techniques such as HE, the PATE, and SMPC have been examined in existing studies, insufficient comparisons have been conducted regarding how these mechanisms perform when applied in different combinations. For instance, no research has ever systematically compared the performance implications of combining HE with the PATE with those implementations that combine HE, the PATE, and SMPC together. Such a comparison analysis gap makes it challenging for practitioners to make informed decisions on the best security configurations for their FL deployments. The evaluation of trade-offs among various sets of security measures, in terms of the computational overhead, communication overhead, model precision, and overall system efficiency, is a relatively less-researched subject compared to some other subjects within the FL field and thus requires additional research.

In this paper, we present a comparative study of different possible combinations of privacy-preserving methods with FL. The analysis included both the performance and security of all the generated models to evaluate the resistance of each model against various attacks. The contributions of this study are as follows:

1. To overcome the challenges of traditional ML, we developed an FL model using an artificial neural network (ANN) based on a Malware Dataset for the detection of malware.
2. To enhance the privacy of FL, we integrated various privacy-preserving techniques with FL. We developed multiple models for the combination of FL with privacy-preserving techniques. The PATE, SMPC, and HE were used for preserving the privacy of the FL model. The main focus of this work was to analyze how different privacy-preserving techniques interact when combined in federated settings, their collective impact on the model performance, and the resulting security guarantees.
3. We evaluated the generated models against various attacks, including poisoning attacks, a model inversion attack, a backdoor attack, and a man in the middle attack.
4. To the best of the authors' knowledge, this is the first work to analyze the performance and security of different possible combinations of privacy-preserving techniques in

FL. The generated models did not reduce the performance by a large percentage, but they improved the models' robustness against various attacks. All combinations performed better than the base FL model for all evaluated attacks.

The rest of this study is arranged as follows: Section 2 addresses the background and related work on FL and its privacy-preserving techniques. In Section 3, the proposed methodology is outlined. Section 4 includes the experimental setup. Section 5 provides the results, including a performance and security analysis. Section 6 provides a discussion. Finally, Section 7 addresses the conclusions and suggested future work.

2. Background and Related Work

2.1. Federated Learning

FL is an ML approach used for distributed learning. The architecture of this approach involves various components, including the central server, clients (devices), the global model, and local models [33]. Each component plays an important role in effective model training and data privacy. The central server is responsible for coordinating the overall FL process, aggregating local model updates, and maintaining the global model while the clients, which represent individual devices participating in the FL process, perform local model training on their local datasets [34]. The global model is the primary training model that is shared and modified by all clients, while local models represent locally updated copies or modifications of the global model created by each client [35].

FL is a decentralized ML approach that enables several devices to collaboratively learn useful and essential information from their owned data without sharing them [36]. It allows for the collaborative training of an ML model through various devices on their local datasets [37]. The clients (devices) perform local training on their data. After local model training, the clients then send their local model updates (not raw data) to the central server [38]. The central server aggregates these model updates from all the clients participating in the FL process [39]. After aggregation, the central server improves the global model by updating it using the aggregated model updates. The updated global model is then sent back to clients to perform further training [40].

Scalability, the need for significant and powerful computational resources and extensive training data, data privacy issues, and possible bottlenecks in the transport and processing of data are some of the major obstacles faced by centralized advanced ML, including ANNs [41]. In order to mitigate these obstacles, FL can be integrated with ANNs. There are various studies that have improved performance and enhanced privacy by integrating FL with neural networks, such as [42–44]. The use of FL is becoming more and more effective in improving malware detection. FL enhances malware detection systems' accuracy and effectively addresses privacy issues by facilitating decentralized data processing.

Compared to traditional ML approaches, FL ensures user data privacy, makes anomaly detection in IoT networks easier, improves malware detection, enhances the model accuracy, and provides better accuracy in the detection of attacks [45]. For Android malware classification, FL enhances the detection of malware and outperforms traditional deep learning techniques in terms of detection accuracy, scalability, and user privacy preservation [46]. In healthcare, through collaborative model training across multiple clients (four hospital networks), FL improves malware detection by increasing accuracy, patient privacy preservation, and resilience against malware threats [47].

2.2. Security in Federated Learning

The FL approach is vulnerable to several types of attacks, such as poisoning attacks, backdoor attacks, model inversion attacks, membership inference attacks, Property Infer-

ence Attacks, and model extraction [48]. Depending on the attack method by which the attacker changes the parameters of the local model to create a poisoned model, poisoning attacks can be classified as either data poisoning attacks or model poisoning attacks [49]. Regarding data poisoning attacks, attackers perform the attack on local training datasets either by manipulating labels (e.g., label flipping) or samples (e.g., adding noise to the training dataset) [50].

For model poisoning, the attacker directly manipulates local model updates (e.g., adding noise to model updates or sending arbitrary model updates) [51]. Poisoning attacks influence the global model's performance [49]. To compromise the global model in an FL configuration, backdoor attacks modify the local models. The attacker's goal in these attacks is to incorporate a trigger into one or more local models so that, when the trigger is present in the data inputs, the global model behaves in a particular way [52]. Backdoor attacks mislead the global model to generate inaccurate outputs when given backdoor inputs.

In model inversion attacks, the attacker tries to extract sensitive information or reconstruct the original training dataset by manipulating the final global model output [53]. In membership inference attacks, the attacker tries to discover if a particular data point is a member or non-member of the training dataset used for model training [54]. In Property Inference Attacks, the attacker aims to extract sensitive features included in the dataset used for model training [54]. For model extraction or stealing attacks, the adversary tries to steal model functionality by stealing the model parameters and hyperparameters. By only querying the target model, an attacker can create a replacement model in model extraction attacks that has nearly all of the functions of the target model [54].

In ML, the attack success rate (ASR) measures how well an attack can breach, manipulate, or compromise an ML model [55]. This success rate can be used as an evaluation metric for the effectiveness of attacks on ML models [56]. This metric is essential for assessing how resilient ML models are to different types of attacks, particularly in light of the growing concerns surrounding adversarial attacks [57]. The metric provides the percentage of successful attacks on the ML model [58,59].

Equation (1) shows the general calculation of the ASR. The ASR can by calculated by dividing the number of attempts in which the attack actually succeeded in fooling the ML model by the total number of attack attempts, and the result is then multiplied by 100 to obtain the percentage. Specifically, in the case of poisoning attacks, the ASR will be the proportion of successful poisoning attempts to the total number of poisoning attempts, as noted in studies of ML threats that highlight the exploitation of vulnerabilities in ML systems [60,61]. For backdoor attacks, it represents the proportion of successful backdoor triggers to the total number of trigger attempts [59,62].

$$\text{ASR} = \frac{\text{Number of Successful Attacks}}{\text{Total Number of Attack Attempts}} \times 100\% \quad (1)$$

For a model inversion attack, the ASR is the proportion of correctly reconstructed inputs to the total number of reconstruction attempts [63,64]. Regarding a membership inference attack, it is the ratio of correctly inferred membership instances to the total number of inference attack attempts [65,66]. For model stealing attacks, it will be the ratio of successfully replicated functionalities to the total number of stealing attack attempts [67,68], and so on for other attacks. This metric is important because model theft can disrupt competitive advantages in ML and result in large financial losses [69].

2.3. Privacy-Preserving Techniques

There are several techniques for privacy preserving in ML, such as the PATE, SMPC, and HE.

2.3.1. Private Aggregation of Teacher Ensembles

The PATE [70] is considered a popular privacy-preserving approach [71]. It is an ML-based framework in which numerous ensembles of teacher models and a student model are combined to create private models [72]. Teacher models are trained on disjoint private datasets [73]. Based on noisy voting from all the teacher models, the student model performs prediction [74]. The process involves data partitioning, teacher model training, the voting of teacher models, noise addition, aggregation, and student model training. In the data partitioning phase, the private training dataset is partitioned into n subsets, as in Equation (2).

$$\mathrm{D(X, Y)} = \{D_1(X_1, Y_1), D_2(X_2, Y_2), D_3(X_3, Y_3), \ldots, D_n(X_n, Y_n)\} \tag{2}$$

where X is the input training data, Y is the corresponding target class labels, D is the training dataset partitioned into n disjoint subsets, n is the total number of partitions inside the dataset, and $Di(X_i, Y_i)$ represents the ith partition including a subset of the training data and its corresponding class labels.

In the teacher model training phase, each teacher model, T, is trained on one of n data subsets that have been divided before. The teacher model voting is performed as in Equation (3). Noise is then added to the predictions (votes) of the teacher models using a Laplace or Gaussian distribution mechanism. After that, the aggregation of these votes is performed using Equation (4).

$$V(X) = \{T_1(X), T_1(X), T_1(X), \ldots, T_n(X)\} \tag{3}$$

where $T_i(X)$ is a vote or prediction for the input X made by the ith teacher model, n represents the total number of teacher models, and $V(X)$ is a vector containing the number of teacher models that were able to predict each potential class.

The student model is then trained using a public dataset which is labeled with the output of the aggregating teachers, Equation (4), as in [75]. There have been several studies that utilized the PATE, such as [14,16,76,77].

$$Y_{PATE}(X) = argmax(V(x) + noise) \tag{4}$$

where $V(x)$ stands for the vector of the teacher model vote counts for each potential output class for input X, noise represents the random noise extracted from a Gaussian or Laplace distribution, $argumax$ chooses the class with the most noisy votes, and $Y_{PATE}(X)$ represents the final aggregated prediction.

2.3.2. Secure Multi-Party Computation

SMPC is sometimes referred to as multi-party computation (MPC) or SMC or privacy-preserving computation [78]. It is a cryptographic technique, enabling several parties to jointly compute the public function $f(x_1, x_2, x_3, \ldots, x_j) \rightarrow (y_1, y_2, y_3, \ldots, y_j)$, where $P_j : x_j \rightarrow y_j$ [79]. It allows these parties to collaborate to compute the result from their private individual data inputs without sharing any of their input data with other participants [80].

Only the result $(Y_1, Y_2, Y_3, \ldots, Y_j)$ and its own data inputs, (X_j), are available to each party [78]. This keeps private data safe, even in the event that one party is compromised, and allows even untrustworthy people to securely collaborate on sensitive data [81]. There have been several studies that utilized SMPC, such as [18,82–84].

2.3.3. Homomorphic Encryption

HE is a cryptographic technique which offers the ability to perform computations on an encrypted form of data called ciphertext without the necessity for decrypting the ciphertext first, ensuring data confidentiality [85]. Cryptographic results are obtained from the computation; moreover, this technique guarantees that the output that is decrypted will match the output that was calculated using the original plaintext dataset [78]. According to several studies, such as [86,87], there are various categories of HE techniques based on the type and number of operations enabled by the system, including partially [88], somewhat (SWHE) [89], and fully homomorphic encryption (FHE) [29].

The most mathematically constrained type of HE is partially homomorphic encryption (PHE), which is also the most computationally practicable [90]. PHE provides either an addition or multiplication operation for an unlimited number of times [91]. SWHE enables both simple addition and multiplication for a limited number of operations [92]. FHE supports any arbitrary operations on encrypted data [78]. Though it requires more processing than any other HE kind, FHE is the strongest [90]. There are various HE schemas, but the most popular [93] are TFHE [94], BFV [95], and CKKS [96].

The CKKS (Cheon–Kim–Kim–Song) schema supports approximate encrypted computations over real and complex numbers, which is useful in ML scenarios [97]. The Brakerski/Fan–Vercauteren (BFV) scheme enables exact computations on integer ciphertexts [98]. The FHE over the Torus (TFHE) scheme allows for efficient Boolean operations on encrypted data and uses a bootstrapping method to control ciphertext noise [99]. There have been several studies utilizing the CKKS schema, including [100–102]. There have also been various studies utilizing the BFV schema, such as [103,104]. There have been multiple other studies utilizing the TFHE schema, such as [105,106].

2.4. Related Work

In [107], the authors proposed an FL framework for malware detection across IOT devices. Their framework utilizes both supervised and unsupervised FL models, including a multi-layer perceptron and autoencoder. They used the N-BaIoT dataset, a dataset that models the network traffic of many actual IoT devices when they are compromised by malware. They performed a Performance Comparison between the federated and centralized approaches. According to their results, privacy can be maintained while achieving performance levels similar to those of centralized models using the federated approach. Supervised FL produced high accuracy above 99%, the same as for central approaches. Unsupervised FL also achieved results similar to the central approach: a 99.98% TPR for both known and new devices, along with a 94.84% TNR (multi-epoch avg) and 95.12% TNR (mini-batch avg) in the case of known devices and 92.61% (multi-epoch avg) and 91.78% TNRs (mini-batch avg) for new devices. They did not apply any privacy-preserving technique besides FL.

In [108], the authors proposed FEDriod, an FL framework for Android malware detection based on a residual neural network (ResNet). They used the CIC, Drebin, and Contagio datasets. Their FEDriod framework achieved a 98.53% F1-score. They did not apply privacy-preserving techniques along with FL. In [109], the authors proposed an FL framework for the detection of ransomware based on RNNs, recurrent neural networks. Synthetic data were utilized for both phases of training and evaluating the model. Their model produced an accuracy of 94.7%, a precision of 92.3%, a recall of 91.8%, an F1-score of 92.0%, and an AUC-ROC of 96.1%. Besides FL, they did not apply privacy-preserving techniques.

In [110], the authors proposed a framework called FedHGCDroid, for the detection and classification of Android malware. They built their HGCDroid model using both a Convolutional Neural Network for malware's statistical features and graph neural networks

(GNNs) for malware's graphical features. They used the Androzoo dataset. For malware detection, the HGCDroid model achieved an accuracy of 91.3%, a precision of 90.8%, a recall of 92.79%, and a 91.29% F1-score. For malware classification, the model produced an accuracy of 83.29%, a precision of 83.45%, a recall of 83.85%, and an 83.67% F1-score. This framework did not integrate privacy-preserving techniques with FL.

In [111], the authors proposed the SIM-FED model for the detection of malware in IOT devices. Their model applies both FL and deep learning. A lightweight one-dimensional CNN with improved hyperparameters is used in the model. The FedAvg strategy is utilized to include the outcomes of local models in the suggested model. The accuracy achieved by the SIM-FED model was 99.522%. In [11], the authors proposed a system called FedML-HE. The system integrated the FL approach with HE for securing the model aggregation process. The PALISADE and TenSEAL libraries were used for implementing HE. They used the CIFAR-100 and wikitext datasets.

During the training phase, the system encrypted only sensitive parameters for updating the local model to reduce the overhead. For ResNet-50, their system achieved a reduction of $\sim 10x$. For BERT, their system could achieve a reduction of up to $\sim 40\times$. This system integrated only one privacy-preserving technique, HE, with FL. The provided system preserved only the model updates. In [32], the authors proposed FLCP, which is an FL-based framework, along with enhanced communication efficiency and privacy preservation. For privacy preservation, they integrated FL with HE and DP. FLCP was evaluated on the MNIST and CIFAR-10 datasets using a unified (CNN) Convolutional Neural Network architecture.

Local updates, weight compression, DP addition, and HE for sending encrypted updates are the processes within their FLCP framework. Each client performs local training and computes local updates. After the computation of the local model updates, clients perform weight compression using the AWC-FedAvg algorithm. Following weight compression, noise is added to the compressed local model updates using a Laplace mechanism. After the addition of DP, HE is applied to the noisy compressed updates. The encrypted updates are finally sent by the clients to the central server, which aggregates them to update the global model. The provided system tries to preserve only local model updates, and a possible loss of information may exist due to compression.

In [112], the authors proposed an FL framework using a Convolutional Neural Network in the field of Sixth-Generation (6G) wireless networks and the Internet of Medical Things (IoMT). The framework integrated SMPC using additive secret sharing for secure aggregation. A breast cancer dataset from the Histopathological Database was used. The ResNet model achieved a training accuracy of 98% and a validation accuracy of 90%. The AlexNet model achieved a training accuracy of 95% and a validation accuracy of 85%. The framework integrated only one privacy-preserving technique, SMPC, with FL. The complexity of implementing SMPC can produce a high overhead. The proposed framework was mainly focused on securely aggregating only the local model updates rather than the raw data themselves.

In [12], the authors proposed an FL framework that combines SMPC and blockchain technology in IoT networks. The framework uses SMPC protocols for the secure aggregation of model updates from participants without disclosing personal client data, while blockchain technology guarantees transaction immutability and transparency. In [23], the authors proposed FL-PATE, which is a differentially private FL framework with knowledge transfer. The framework integrates the PATE with FL to enhance privacy. In [24], the authors proposed the FREDY framework based on Convolutional Neural Networks. FREDY integrated FL, the PATE, and knowledge transfer.

The PATE was integrated into FREDY through the federated training of several teacher models, inference on publicly available unlabeled data, the aggregation of predictions with Laplace noise, and the training of a student model on the labeled data obtained from this process. The CIFAR10 and MNIST datasets were used. The study showed that for both the MNIST and CIFAR-10 datasets, increasing the number of clients and the privacy parameter ε generally improved the test accuracy of the student model. MNIST models performed better than CIFAR-10 models. Using MNIST, the 25-client model achieved the maximum accuracy of $\sim$99% at $\varepsilon = 1$.

Using CIFAR-10, the 25-client FL model (FREDY) attained the maximum accuracy of $\sim$79% when $\varepsilon = 1$. When $\varepsilon = 0.2$, FREDY outperformed the baseline model in terms of its membership inference attack performance, reducing all metrics with a $\sim$25% drop. The framework was evaluated against only one attack. It also integrated only one privacy technique with FL. In [28], the authors presented a framework based on FL and HE for intrusion detection in the Internet of Vehicles (IoVs) with limited computing resources. They used a Deep Neural Network (DNN) and the Edge-IIoT dataset. The framework includes two processes, pre-learning and privacy-preserving learning.

The proposed framework integrates the approaches of both local and centralized learning. Initially, Vehicle Users partition their data by determining the optimal amount of local training and offloading data based on their computational resources. Using HE, Vehicle Users encrypt offloaded data before transmitting it to the centralized server (CS) via Roadside Units. The central server compiles the received ciphertext into an encrypted dataset and creates its own server model for training on this ciphertext dataset. Concurrently, Vehicle Users perform local training using the server's distributed global model, which is initialized from the server model. Following local training, Vehicle Users encrypt and transmit their local model updates to the central server for aggregation.

The central server redistributes the updated global model to Vehicle Users for further training. The presented framework achieved a high accuracy of approximately 91% in attack detection. The framework was not evaluated against any attack. It also integrates only one privacy technique with FL. In [25], the authors presented an FL-based framework using Convolutional Neural Networks called PrivatEyes, to protect gaze estimation data. The framework integrated SMPC using secret sharing. The MPIIFaceGaze, GazeCapture, and NVGaze datasets were used for evaluation. The mean angular errors (MAEs) were 6.3 for MPIIGaze, 6.2 for MPIIFaceGaze, and 0.8 for NVGaze. Using MPIIGaze, PrivatEyes produced a Re-identified of (0/15) and visual score of 4%.

The accuracy for the MPIIFaceGaze dataset dropped from 40% in the first round to 33% in the final round of the evaluation of membership inference attacks, while the accuracy for the MPIIGaze dataset dropped from 33% to 13%. Likewise, across the rounds, the NVGaze dataset revealed a drop from 37% to 12%. Furthermore, the prediction accuracy for gender categorization was significantly reduced by attribute inference attacks across the datasets; MPIIFaceGaze showed a decline from 5 out of 15 participants in the first round to 2 out of 15 in the last round, MPIIGaze showed a decline from 3 out of 15 participants in the first round to 0 out of 15 in the last round, and NVGaze showed a decline from 10 out of 32 participants in the first round to 1 out of 32 in the last round.

In [26], the authors proposed an FL system for the Internet of Medical Things. The system integrates SMPC and an additive secret-sharing method, along with FL. With this method, the federated training model's gradient parameters—rather than the actual data—are protected. The prediction of the suggested model is made using a Convolutional Neural Network, which uses a weight-sharing technique. The model's weight is encrypted through additive secret sharing and multiple computations. The accuracy rate of the suggested model is 97%. In addition, the suggested model's F1-score is 89%. To the best

of the authors' knowledge, this is the first study comparing different possible combinations of privacy-preserving techniques with FL. There also is not previous work considering the combination of FL with more than two privacy-preserving methods out of the PATE, SMPC, and HE.

3. Methodology

3.1. Federated Learning Model

This research provides multiple new contributions to the area of privacy-preserving machine learning. To start, we propose a holistic framework that is the first to integrate federated learning (FL) with a wide range of privacy-preserving methods, extending beyond typical dual-method solutions. We provide a systematic evaluation of diverse sets of privacy-preserving methods paired with FL, examined across multiple dimensions. Specifically, we conducted extensive experiments to measure the resilience to multiple types of privacy attacks, tallying success rates under a variety of attack scenarios. We further examined the practical implications of these pairings on a number of performance indicators, including the model accuracy, computational overhead, and training duration efficiency. This thorough testing was particularly critical, as certain privacy-preserving methods, e.g., homomorphic encryption (HE), involve significant computational complexity. Our extensive analysis offers useful insights into the privacy protection, model performance, and computational cost trade-offs, thus serving as a helpful guideline for the deployment of privacy-preserving federated learning systems.

The proposed FL system uses a deep feed-forward ANN architecture that is intended for the classification of multiple classes. To validate our privacy-preserving FL framework, we implemented a weight-dependent model architecture. We selected an artificial neural network (ANN) for this purpose, as illustrated in Figure 1. Figure 1 illustrates the architecture of the designed and used ANN classifier. Unlike a common ANN, which uses a single ANN model, the FL system has a generic ANN model and a number of local ANN models, and all have the same architecture. The architecture of the used ANN classifier is defined on the server side. This ANN classifier is implemented using the Keras Sequential API, which enables a linear stack of layers. It consists of a sequential model with multiple layers. In particular, it consists of five conceptual layers: an input layer, three hidden layers, and an output layer. In detail, the ANN architecture includes eleven individual layers: an input layer, an output layer, and three hidden layers, each including three individual layers.

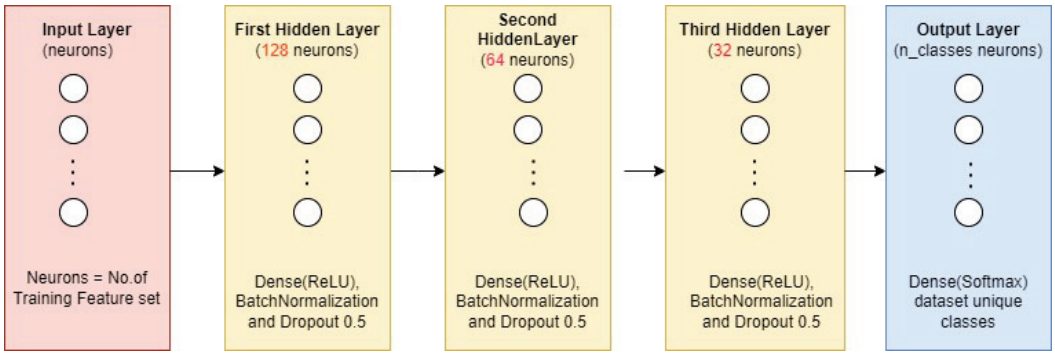

Figure 1. ANN architecture.

The input layer is determined by the training dataset's feature set shape, with the number of neurons matching the feature count. This dynamic design adapts the ANN to various datasets. The model includes three hidden layers using ReLU activation. The first hidden layer has 128 neurons, Batch Normalization, and dropout (0.5). The second and third layers have 64 and 32 neurons, respectively, with similar configurations. The output layer

uses a dense layer with a number of neurons equal to the unique class count, and Softmax activation generates probability distributions. The model is compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy metric.

In FL, a client–server architecture involves a central server and three clients. Clients train locally on subsets, while the server coordinates learning and aggregates models. Initialization sets up a Flask app for communication, using ngrok for a public URL. The server loads, pre-processes, and splits data into training (80%) and testing (20%) sets, applies feature scaling, one-hot encodes target labels, and balances class weights. Each client similarly pre-processes its data.

The FL process begins with model distribution. Clients request the global model, train locally for five epochs, and send updated weights to the server, which aggregates them via averaging. This cycle repeats for five rounds, refining the global model with diverse data. After the FL rounds, the server fine-tunes the global model with early stopping and learning rate reduction, evaluates it using metrics like the accuracy and F1-score, and saves it in an h5 file for future use.

3.2. Implementation of Privacy-Preserving Techniques

Privacy-preserving approaches are crucial in the real-world applications of FL because of the several points of vulnerability present in the FL system. FL disperses computations and data among several devices as opposed to centralizing them. By keeping raw data local, this helps preserve privacy, but exchanging information—specifically, model updates—is still part of the process. During the FL process, local models or the model updates of weights are sent between clients and the central server, and the global model is created through an aggregation process. The communication between FL-participating clients and the central server, as well as the shared model updates, provide possible gaps where privacy may be violated.

Attackers might try to intercept and manipulate the transmitted data between the clients and server. They also might try to analyze and manipulate the final global model to expose sensitive information about the data used for training. Also, if the server is unreliable, it could potentially misuse the data it obtains for performing potential attacks. Malicious clients might try to manipulate their local raw data or manipulate the local model weights themselves before sending them to the server. Furthermore, the aggregation process itself might allow for attacks. These discussed potential real risks in the FL pipeline, which ranges from local model training to data transmission, the aggregation process, and the final global model, highlight the vital necessity for strong privacy-preserving mechanisms to be incorporated along with FL.

3.2.1. Private Aggregation of Teacher Ensembles

In this FL system, the PATE is mainly implemented on the server side. After FL, the system applies the PATE as a post-processing step to provide an extra layer of privacy protection to the FL process. Following the completion of FL rounds, the server starts the PATE procedure by creating a set of multiple teacher models. In this implementation, 10 teacher models are created using the ANN architecture. The amount of information that any one teacher model may disclose about specific data points is limited by ensuring that each teacher model has only been exposed to a small portion of the entire dataset.

Each of these teacher models is then trained using a distinct and randomly selected subset of the server's training dataset. After completing the training of the teacher models, the system begins the aggregation phase. In this aggregation phase, all teacher models' predictions are gathered. In order to mitigate the danger of the leakage of sensitive data, a privacy-preserving aggregation mechanism is employed instead of using raw predictions

from the teacher models. This is accomplished by adding DP using Laplace noise to the total predictions from all of the teacher models. In order to control the noise injection, the privacy budget, which is represented by the epsilon parameter, is set to a smaller value of 0.1 to ensure stronger privacy guarantees.

To determine the final predictions following the noisy aggregation, the class with the highest noisy count for each data point is chosen. The final stage of the PATE process involves the creation and training of the student model. The final global model, which has already undergone the FL process, is used as the student model. The student model is trained on the new aggregated, privacy-preserving labels of predictions generated from the teacher ensemble through the PATE process.

In conclusion, the PATE is implemented as an additional processing phase subsequently to FL. It starts by creating an ensemble of teacher models. It then aggregates these teacher models' predictions using a privacy-preserving aggregation technique employing DP. Finally, the student model (the global model) is subsequently trained using these privacy-preserving predictions. By implementing the PATE in this manner, the advantages of FL (maintaining decentralized data) can be combined with the PATE privacy guarantees.

3.2.2. Secure Multi-Party Computation

In this experiment, SMPC was added to the implementation of the predefined FL system described in Section 3.1 on both the server and client sides. L2 regularization was added to the ANN model. Each dense layer except the output layer has an L2 regularization parameter of 0.01. The PySyft library, which offers tools for privacy-preserving ML, is used for the implementation of SMPC in this FL system. PySyft workers are initialized by both the FL server and clients for secure computations and the aggregation of model weights. The server initializes and creates its own PySyft worker on the server side, whereas each client participating in the FL process also initializes and creates its own PySyft worker.

Specifically, one worker is used for the server, and there are three client workers (one worker for each client). These PySyft workers are responsible for the secure processing and computation of the data and ANN model parameters for the SMPC technique. PySyft tensors are used to manage and transmit the ANN model weights between FL-participating clients and the central server. These tensors are used primarily on the server side in the process of receiving and aggregating updated ANN model weights from clients. When the server receives updated model weights from each client, it transforms them into PySyft tensors for the secure handling of these weights in the subsequent operations of aggregation and DP.

After receiving the updated ANN model weights from all participating clients, the server performs the FL aggregation process on these weights. As the averaged model weights produced by the aggregation process could potentially reveal information about the data of individual clients, Gaussian noise is added to the aggregated average model weights in order to improve privacy and achieve DP. The noise scale is calculated based on the sensitivity and epsilon parameters. The epsilon parameter is set to a smaller value of 0.1 to ensure stronger privacy guarantees, while the sensitivity parameter, which represents the maximum effect on the output that a single data point can have, is also set to 0.1. Using these two parameters, random noise from a Gaussian distribution is generated, scaled properly, and added to the averaged weight.

On the client side, clients train their local models and then transform the resulting weights into PySyft tensors for secure aggregation and manipulation without disclosing the underlying data. All clients ensure the consistent use of L2 regularization. After receiving the global model from the server, each client dynamically applies L2 regularization (0.01) to all applicable layers of the received ANN global model. By performing this regularization

method on the client side, it guarantees Consistency Across Clients by ensuring that all clients apply regularization at the same level, independently of how the model was initially established on the server side. We also tested this implementation of SMPC with FL without the addition of regularization and DP.

3.2.3. Homomorphic Encryption

HE was added to the implementation of the predefined FL system described in Section 3.1 on both the server and client sides. The CKKS (Cheon–Kim–Kim–Song) scheme is used for the implementation of HE. The TenSEAL library is used by both the server and the client to set up an identical context for the CKKS schema. This context configures crucial parameters such as the polynomial modulus degree and coefficient modulus bit sizes, which specify the computational capabilities and security level of the encryption. The polynomial modulus degree is set to 8192, while the coefficient modulus bit sizes are set to [60, 40, 40, and 60]. Additionally, the global scale, which affects the precision of encrypted computations, is set to 2^{40}.

The encryption process is primarily conducted on the client side. The clients encrypt their model weights before sending them to the server. The encryption process can deal with various types of model layers. It behaves differently for each type. It uses the CKKS encryption technique to manage dense layer weights and biases. Each row is encrypted independently as a vector for dense layers, which normally include a 2D array of weights. Bias layers, being 1D, are encrypted as a single vector. These encrypted weights are then prepared for network transmission by serializing them into Base64-encoded strings following encryption.

On the server side, the server performs the aggregation and processing of the encrypted gradients from all participants. When the server receives the clients' encrypted gradients, it first deserializes these received encrypted gradients back into their form of encrypted vectors and then performs the aggregation process on them without decryption. The aggregation process computes the average of these encrypted gradients from all participants. Since the ANN model includes multiple layers, each with a unique set of parameters (weights and biases), the aggregation process is performed layer by layer.

The aggregation is performed separately for each layer in order to maintain the ANN model's structure. Dense layer weights and biases are handled separately. Regarding dense layers (2D), the gradients of each neuron are aggregated individually. All of the clients' encrypted gradients for each neuron are deserialized, and homomorphic addition is used to add them. The outcome is then scaled using homomorphic multiplication by dividing it by the total number of clients. For bias layers (1D), since there is only one vector per layer, the procedure is comparable but less complicated. The bias's client gradients are all deserialized and added together. After that, the outcome is divided by the total number of clients.

Throughout this aggregation process, the data remain in their encrypted form. The homomorphic properties of the CKKS scheme allow for these arithmetic operations (addition and multiplication by a scalar) to be performed on the encrypted data without the need for decryption. After finishing the aggregation, the global model is updated using its result, which is a set of averaged, encrypted gradients that show the overall update from all clients. Specifically, the resulting aggregated gradients are first decrypted by the server. To maintain training stability and enhance privacy, the server then applies gradient clipping, a method that restricts the magnitude of gradients, after decryption.

The process of gradient clipping involves computing the total L2 norm of all combined gradients and comparing it to a maximum norm that has been predetermined, in this case 1.0. In the event that the overall norm is higher than this cutoff, all gradients are

resized proportionately to make sure their total magnitude stays below the threshold. This has dual purposes: it prevents the model from receiving extreme updates that could destabilize the training process, and it enhances privacy by limiting the impact of any single participant's update on the global model. The server updates the global model using these clipped gradients.

HE is integrated with the FL model in the stage of client-side gradient encryption. Clients use the CKKS schema for HE to encrypt their computed gradients before submitting updates to the server. HE is also integrated with the FL model in the stage of server-side secure aggregation performed on encrypted gradients. This is yet another crucial phase in which HE is included. Using homomorphic operations, the server performs aggregation on the encrypted gradients from each client. This makes it possible to perform computations with encrypted material without first decrypting it. In particular, the homomorphic operations used are addition and multiplication. HE is also used in the server-side decryption process. After aggregation, the server decrypts the result to apply the clipping technique in order to maintain training stability and enhance security.

3.3. Configuration Combinations

Table 1 illustrates different combinations of FL with privacy-preserving techniques. The first column shows the model number, and the second column shows the model itself, while the included methods are represented in the last four columns (third, fourth, fifth, and sixth columns). FL, the PATE, SMPC, and HE represent the included methods. The ✓symbol means that this method is included in the corresponding model, while the x symbol means that the method is not included in the model. For example, the third model, FL_SMPC_DP, includes FL and SMPC, and the sixth model, FL_CKKS, includes FL and HE, while the last model, FL_PATE_CKKS_SMPC, applies all the included methods of FL, the PATE, SMPC, and HE. The main purpose of including several models was to conduct a comparative study with an analysis of the performance and security of different possible integrations of privacy-preserving methods with FL.

Table 1. Combined privacy preserving techniques.

		Included Methods			
Model No.	**Model**	**FL**	**PATE**	**SMPC**	**HE**
1	FL_only	✓	x	x	x
2	FL_SMPC	✓	x	✓	x
3	FL_SMPC_DP	✓	x	✓	x
4	FL_PATE	✓	✓	x	x
5	FL_PATE_SMPC	✓	✓	✓	x
6	FL_CKKS	✓	x	x	✓
7	FL_CKKS_DP	✓	x	x	✓
8	FL_CKKS_SMPC	✓	x	✓	✓
9	FL_PATE_CKKS	✓	✓	x	✓
10	FL_PATE_CKKS_SMPC	✓	✓	✓	✓

The first model, FL_only, represents the base model to which the rest of the nine models were compared. The FL_only model uses a deep feed-forward ANN architecture which consists of five conceptual layers: an input layer, three hidden layers, and an output layer. To implement FL, a client–server architecture involving a central server and three clients is used. Clients train locally on their local data, while the server coordinates learning

and aggregates models. Specifically, the FL process begins with the distribution of the ANN model. Clients request the global model, train locally for five epochs, and send updated weights to the server, which aggregates them via averaging. This cycle repeats for five rounds, refining the global model with diverse data. After the FL rounds, the server fine-tunes the global model with early stopping and learning rate reduction and evaluates it using metrics such as the accuracy, precision, recall, and F1-score.

The second model, FL_SMPC, combines SMPC with FL. The purpose behind the addition of SMPC to FL is to protect the model updates before transmission, which in turn helps in securing the communication between the server and clients. SMPC was added to the implementation of the FL_only model on both the server and client sides using the PySyft library. PySyft tensors are used to manage and transmit weights among clients and the central server. These tensors are used for both clients and the server. On the server side, the tensors are used primarily in the process of receiving and aggregating updated ANN model weights from clients. When the server receives updated model weights from each client, it transforms them into PySyft tensors for the secure handling of these weights in the averaging aggregation process. On the client side, clients train their local models and then transform the resulting weights into PySyft tensors before transmission.

The third model, FL_SMPC_DP, integrates SMPC with FL. This model protects the model updates using SMPC. To provide an extra layer of privacy protection to the FL_only model, SMPC was added to its implementation on both the server and client sides using the PySyft library. PySyft tensors are used to manage and transmit model updates between FL-participating clients and the central server. These tensors are used primarily on the server side in the process of receiving and aggregating updated ANN model weights from clients. As the averaged model weights produced by the aggregation process could potentially reveal information about the data of individual clients, Gaussian noise is added to the aggregated average model weights in order to improve privacy and achieve DP. On the client side, clients train their local models and then transform the resulting weights into PySyft tensors before transmission. To enhance security and generalization, L2 regularization was added to the ANN model on the server side. All clients also ensure the consistent use of L2 regularization. After receiving the global model from the server, each client dynamically applies L2 regularization to all applicable layers of the received ANN global model. The main difference between the FL_SMPC model and FL_SMPC_DP model is that the FL_SMPC model does not include DP and regularization, while the FL_SMPC_DP model includes this extra layer.

The fourth model, FL_PATE, integrates the PATE with FL. To provide an extra layer of privacy protection to the FL_only model, the PATE is implemented as a post-processing step. After the completion of the FL process, the PATE is implemented to protect the final FL model from being attacked. The PATE is mainly implemented on the server side; after the completion of the FL rounds, the server starts the PATE procedure by creating a set of 10 teacher models. Each of these teacher models is then trained using a distinct and randomly selected subset of the server's training dataset. In order to mitigate the danger of the leakage of sensitive data, DP using Laplace noise is added to the total predictions from all of the teacher models. Finally, the student model (the global model) is subsequently trained using the new aggregated, privacy-preserving labels of predictions generated from the teacher ensemble through the PATE process. The student model then serves as the final global model and is used for the evaluation process.

The fifth model, FL_PATE_SMPC, combines SMPC and the PATE with FL. This model protects both the model updates using SMPC and the final FL model through the PATE. SMPC was added to the implementation of the FL_only model and SMPC was added to the FL_only model's implementation on both the server and client sides using the

PySyft library. As the averaged model weights produced by the aggregation process could potentially reveal information about the data of individual clients, Gaussian noise is added to the aggregated average model weights in order to improve privacy and achieve DP. On the client side, clients train their local models and then transform the resulting weights into PySyft tensors before transmission. To enhance security and generalization, L2 regularization was added to the ANN model on the server side. After receiving the global model from the server, each client dynamically applies L2 regularization to all applicable layers of the received ANN global model. After the completion of the FL process with SMPC, the PATE is implemented to secure the final FL model from various attacks. The PATE is mainly implemented on the server side. After the completion of the FL rounds, the server starts the PATE procedure by creating a set of 10 teacher models. Each of these teacher models is then trained using a distinct and randomly selected subset of the server's training dataset. In order to mitigate the danger of the leakage of sensitive data, DP using Laplace noise is added to the total predictions from all of the teacher models. Finally, the student model (the global model) is subsequently trained using the new aggregated, privacy-preserving labels of predictions generated from the teacher ensemble through the PATE process. The student model then serves as the final global model and is used for the evaluation process.

The sixth model, FL_CKKS, integrates HE with FL. By adding HE to the FL_only model, the data and model updates are protected during computation and transmission, which in turn helps in securing the communication between the server and clients. Using the CKKS schema, HE was added to the implementation of the FL_only model on both the sides of the server and clients. Clients encrypt their model updates before sending them to the server, while the server performs the aggregation and processing of the encrypted updates from all participants without the need for decryption. The aggregation process computes the average of these encrypted updates from all participating clients. Throughout this aggregation process, the data remain in their encrypted form without the need for decryption. After the completion of the aggregation process, the global model is updated using its result, which is a set of averaged, encrypted gradients that show the overall update from all clients.

The seventh model, FL_CKKS_DP, integrates HE (using CKKS schema) and DP with FL. Through HE, the data and model updates are protected during computation and transmission, which in turn helps in securing the communication between the server and clients. HE based on the CKKS schema was added to the implementation of the FL_only model on both the sides of the server and clients. Clients encrypt their model updates before sending them to the server, while the server performs the aggregation and processing of the encrypted updates from all participants without the need for decryption. The aggregation process computes the average of these encrypted updates from all participating clients. Throughout this aggregation process, the data remain in their encrypted form without the need for decryption. To preserve the final model's accuracy, as well as increase the model's robustness against attacks targeting individual client's training data, DP is integrated with HE and FL. DP is used to protect the privacy of individual client's training data. The final global model is updated with noisy model updates. Using the Gaussian distribution mechanism, DP is added to the aggregated model updates of all FL-participating clients.

The eighth model, FL_CKKS_SMPC, integrates HE (using the CKKS schema) and SMPC with FL. Through HE, the FL_CKKS_SMPC model protects the data and model updates during computation and transmission, which in turn helps in securing the communication between the server and clients. Using the CKKS schema, HE was added to the implementation of the FL_only model on both the sides of the server and clients. Clients encrypt their model updates before sending them to the server, while the server

performs the aggregation and processing of the encrypted updates from all participants without the need for decryption. The aggregation process computes the average of these encrypted updates from all participating clients. Throughout this aggregation process, the data remain in their encrypted form without the need for decryption. Through SMPC, the FL_CKKS_SMPC model protects the aggregation process itself. SMPC is implemented during the aggregation process. The aggregation process is performed using multiple pysyft workers and distributes data to them. Specifically, a list of workers is created, one for each encrypted gradient set. The encrypted gradients are then distributed to the respective workers. The aggregation is then performed across all workers.

The ninth model, FL_PATE_CKKS, combines the PATE and HE (using the CKKS schema) with FL. This model protects both the final model through the PATE and the model updates through HE. Using the CKKS schema, HE was added to the implementation of the FL_only model on both the sides of the server and clients. Clients encrypt their model updates before sending them to the server, while the server performs the aggregation and processing of the encrypted updates from all participants without the need for decryption. The aggregation process computes the average of these encrypted updates from all participating clients. Throughout this aggregation process, the data remain in their encrypted form without the need for decryption. After the completion of the FL process with HE, the PATE is implemented to protect the final global model from being attacked. The PATE is mainly implemented on the server side; after the completion of the FL rounds, the server starts the PATE procedure by creating a set of 10 teacher models. Each of these teacher models is then trained using a distinct and randomly selected subset of the server's training dataset. In order to mitigate the danger of the leakage of sensitive data, DP using Laplace noise is added to the total predictions from all of the teacher models. Finally, the student model (the global model) is subsequently trained using the new aggregated, privacy-preserving labels of predictions generated from the teacher ensemble through the PATE process. The student model then serves as the final global model and is used for the evaluation process.

The tenth model, FL_PATE_CKKS_SMPC, combines the PATE, HE, and SMPC with FL. This model protects the final model through the PATE, the model updates through HE, and the aggregation process through SMPC. HE was added to the implementation of the FL_only model on both the sides of the server and clients. Clients encrypt their model updates before sending them to the server, while the server performs the aggregation and processing of the encrypted updates from all participants without the need for decryption. The aggregation process computes the average of these encrypted updates from all participating clients. Throughout this aggregation process, the data remain in their encrypted form without the need for decryption. Through SMPC, the FL_CKKS_SMPC model protects the aggregation process itself. SMPC is implemented during the aggregation process. The aggregation process is performed using multiple pysyft workers and distributes data to them. Specifically, a list of workers is created, one for each encrypted gradient set. The encrypted gradients are then distributed to the respective workers. The aggregation is then performed across all workers. After the completion of the FL process with HE and SMPC, the PATE is implemented to protect the final global model from being attacked. The PATE is mainly implemented on the server side. After the completion of the FL rounds, the server starts the PATE procedure by creating a set of 10 teacher models. Each of these teacher models is then trained using a distinct and randomly selected subset of the server's training dataset. In order to mitigate the danger of the leakage of sensitive data, DP using Laplace noise is added to the total predictions from all of the teacher models. Finally, the student model (the global model) is subsequently trained using the new aggregated, privacy-preserving labels of predictions generated from the teacher ensemble through the

PATE process. The student model then serves as the final global model and is used for the evaluation process.

In the FL_CKKS_DP model, using the Gaussian distribution mechanism, DP is added to the aggregated gradients of all FL-participating clients. The amount of DP noise introduced depends on a number of significant elements. The two most important are Delta (δ) and epsilon (ε). As detailed in Table 2, various privacy budgets, ε, from extremely strong privacy (0.01) to moderate privacy (5.0), were analyzed to understand the trade-off between the privacy budget ε and the utility of the model. Although raising the value of ε might improve accuracy, it is only appropriate for applications with weaker privacy assurances; therefore, for privacy-preserving applications, we should choose strong privacy values. Despite the larger values of ε yielding superior accuracy of up to 88.40% at $\varepsilon = 5.0$, we opted for $\varepsilon = 0.1$ due to it providing robust privacy guarantees while still ensuring a satisfactory accuracy of 70.50%. For the subsequent analysis of attacks, we focused on $\varepsilon = 0.1$ to evaluate security under strict privacy conditions. Delta δ (the probability of privacy violation) was set to 1×10^{-5} to guarantee a negligible risk of privacy failure. The amount of the output that could be varied by adding or removing a single data point is called the sensitivity. Because the gradients were normalized, the sensitivity parameter was set to 1.0.

Table 2. The impact of DP on the final model's accuracy.

Privacy Budget (ε)	FL_CKKS_DP Accuracy
0.01 (extremely strong privacy)	60.10%
0.1 (very strong privacy)	70.50%
0.5 (strong privacy)	75.90%
1.0 (strong privacy)	84.10%
2.0 (moderate privacy)	85.80%
5.0 (moderate privacy)	88.40%

3.4. Evaluation Metrics

Besides the time, accuracy, precision, recall, F1-score, and classification report, the ASR was used for evaluation. The time metric represents the total training time, which is the overall duration from the beginning of the FL process to its completion, including all rounds of the FL process along with the fitting of the final model. The ASR is the proportion of successful attack attempts to the total number of attack attempts. The percentage of accurate predictions—both true positives and true negatives—among all instances analyzed is known as the accuracy.

The precision is the ratio of true positive predictions to all positive predictions (true positives + false positives), which is the positive predictive value of the model. The recall, or sensitivity, is calculated as the ratio of true positive predictions to all actual positive cases (true positives + false negatives), which reflects the ability of the model to detect positive cases comprehensively. The F1-score is a single score that achieves a balance between the precision and recall by taking the harmonic mean of both metrics. The main metrics for each class in the classification problem are comprehensively summarized in a classification report. The precision, recall, and F1-score for every class are usually included in a classification report. The report also provides the model's overall accuracy, along with the macro average and weighted average for every class.

4. Experimental Setup

4.1. Dataset

The experiments were conducted on the Malware Dataset from Kaggle https://www.kaggle.com/datasets/blackarcher/malware-dataset, which is a classification-based Portable Executable (PE) dataset on malicious and benign files. It was created with the aid of a Python 3.11 library and includes the data of both normal and malware PE files. It can be used for the testing and training of various ML models. The main purpose of this dataset is for ML and malware identification.

The dataset consists of a total of 100,000 files. Each file represents an entry in the dataset. The dataset is balanced as it contains 50,000 samples for malware and 50,000 samples for benign files. The dataset consists of a total of 35 feature columns, such as the hash, which is a unique identifier for each file, and the classification, which represents the labeled feature in the dataset as "benign" or "malware".

Specifically, the dataset provides a dynamic examination of Android apps and includes 35 features (dimensions) and 100,000 samples for examining runtime behavior. Each of 100 distinct programs (50 malicious and 50 benign) that were monitored for 1000 ms provided the data. The features for identification include the classification (object/string type), which classifies each sample as either "malware" or "benign", and the hash (object/string type), which contains either SHA-256 hashes or APK filenames that uniquely identify each application. The remaining 33 features are integers (int64). Table 3 shows the feature set of the dataset.

Table 3. Malware dataset features.

Feature	Feature Explanation	Feature Type
hash	Unique identifier for each file	Identification and Classification
classification	Indicates whether the entry file is classified as "malware" or "benign"	Identification and Classification
millisecond	Time offset within each file's time series data	Identification and Classification
state	Current state of the process	Process State and Priority
prio	Priority value	Process State and Priority
static_prio	Static priority	Process State and Priority
normal_prio	Normal priority	Process State and Priority
policy	Scheduling policy	Process State and Priority
vm_pgoff	Virtual memory page offset	Memory Usage and Management
vm_truncate_count	Virtual memory truncated count	Memory Usage and Management
task_size	Size of the task	Memory Usage and Management
cached_hole_size	Size of cached memory hole	Memory Usage and Management
free_area_cache	Free area cache size	Memory Usage and Management
mm_users	Memory management users	Memory Usage and Management
map_count	Number of memory mappings	Memory Usage and Management
hiwater_rss	High-water mark for resident set size	Memory Usage and Management
total_vm	Total virtual memory	Memory Usage and Management
shared_vm	Shared virtual memory	Memory Usage and Management
exec_vm	Executable virtual memory	Memory Usage and Management

Table 3. *Cont.*

Feature	Feature Explanation	Feature Type
reserved_vm	Reserved virtual memory	Memory Usage and Management
nr_ptes	Number of page table entries	Memory Usage and Management
end_data	End of data segment	Memory Usage and Management
last_interval	Last scheduling interval	CPU Usage and Scheduling
nvcsw	Number of voluntary context switches	CPU Usage and Scheduling
nivcsw	Number of involuntary context switches	CPU Usage and Scheduling
utime	User mode time	CPU Usage and Scheduling
stime	System time	CPU Usage and Scheduling
gtime	Guest time	CPU Usage and Scheduling
cgtime	Cumulative guest time	CPU Usage and Scheduling
signal_nvcsw	Signal-related voluntary context switches	CPU Usage and Scheduling
fs_excl_counter	File system exclusive counter	File System and I/O
min_flt	Minor page faults	File System and I/O
maj_flt	Major page faults	File System and I/O
usage_counter	Usage counter	Miscellaneous
lock	Lock value	Miscellaneous

4.2. Hardware and Software Environment

The experiments were performed on Google Colab. The CPU specification was an Intel Xeon CPU @ 2.20 GHz (two cores, four threads, x86_64 architecture). The total memory was 12 GB. The Python programming language was used for all implementations. For testing and training the ML models, a reduced dataset was generated, which was a representative subset of the original Malware Dataset offered by Kaggle with the same balance of classification feature labels.

The reduced dataset consisted of 5000 malware and benign files extracted from the original dataset. This reduced dataset decreased the size of the original full dataset while maintaining the same class distribution in the classification column. In all experiments, 20% of the data were used for testing and 80% for training the ML models.

4.3. Attack Simulation

Due to the wide scope of our study, which took into consideration a range of privacy-preserving techniques (PATE, SMPC, HE) and their combinations, we had to compare the models under different types of attacks. Each privacy-preserving technique was developed to guard against certain vulnerabilities, thus making it essential to compare them on several attack vectors. The arena for our comparison encompassed poisoning attacks, model inversion attacks, backdoor attacks, and man in the middle (MITM) attacks. Although every kind of attack inherently demands a particular definition and measurement method, this multi-faceted test strategy offered a better insight into the protection capabilities of every privacy-maintaining configuration.

4.3.1. Poisoning Attack

This experiment simulated a realistic attack scenario in which a malicious client participating in the FL system tries to perform an attack by manipulating the global model through poisoning its data, and the central FL server must identify and mitigate the

effect of these attacks. All models were evaluated against this attack. The attack was implemented on the client side, specifically in the malicious client, which was the first client. The malicious client altered its training data before training the local model so that the data were changed for the malicious client prior to training.

For an untargeted poisoning attack, the data modification stage was performed by randomly flipping some portion of the classification labels to the wrong classes. For a targeted poisoning attack, the data modification stage was performed by randomly flipping some portion of the classification labels to the target class. A flipping ratio of 0.3 was used. After modifying these labels, the malicious client performed local model training on its poisoned data. The poisoned model weights were then sent to the server by the infected client. The server aggregated the model weights updates from all clients, so the malicious client's poisoned weights were also aggregated with the honest clients' weights.

Following the completion of the FL process, the attack was evaluated on the server side. The evaluation quantified the extent to which the poisoning attack was successful in altering and affecting the performance of the model. The ASR calculated the overall misclassification rate introduced by the attack as the proportion of incorrect misclassifications after the attack. Equation (5) shows how the poisoning attack success rate was measured for an untargeted attack, while Equation (6) calculates the targeted poisoning attack success rate. In order to evaluate the overall performance of the model, the server additionally computed common metrics such as the accuracy, precision, recall, and F1-score.

$$UPASR = \frac{MI}{I} \tag{5}$$

where $UPSR$ refers to the untargeted poisoning ASR, MI is the total number of misclassified instances, and I is the total number of instances.

$$TPASR = \frac{TMI}{INT} \tag{6}$$

where $TPASR$ stands for the targeted poisoning ASR, TMI is the number of instances successful misclassified as the target class, and INT is the total number of instances not originally in the target class.

4.3.2. Backdoor Attack

This experiment simulated a realistic attack scenario in which a malicious client participating in the FL system tries to perform a backdoor attack by manipulating the global model through manipulating and altering its local data, and the central FL server must identify and mitigate the effect of these attacks. All models were evaluated against this attack. The backdoor attack was implemented on the client side, specifically in the malicious client, which was specified with a backdoor client id. The first client was the malicious client for all experiments. The malicious client performed the attack by manipulating its local dataset before training.

Specifically, the malicious client altered 10% of its local training dataset by modifying the first dataset feature to the trigger value and also altering the related labels to those of a target class (the last class in the dataset). For model training, every client—even the malicious one—performed training on its own dataset and then sent model updates to the server, so when the malicious client was participating in the FL process, it updated the server with its local model updates, which had been attacked. The backdoor input was unintentionally included by the server when it aggregated the model updates from all clients, including the malicious one.

Through several FL rounds, the attack continued, progressively integrating the backdoor input into the global model. Once the FL process had finished, the attack evaluation

was performed on the server side. The proportion of the backdoor input samples that a model identifies as the target class is known as the ASR. A higher ASR indicates a more successful attack. Equation (7) shows how the backdoor attack success rate was calculated.

$$BASR = \frac{CBS}{BS} \tag{7}$$

where $BASR$ stands for the backdoor ASR, CBS is the number of correctly classified backdoor samples, and BS is the total number of backdoored samples.

4.3.3. Model Inversion Attack

In this experiment, an attack was conducted on the server script, specifically on the server's final global model after the completion of the FL rounds. Finding an input that the model identifies as belonging to the target class is the aim of this attack, as it may disclose confidential information about that class. Using the trained global model, the attack tries to reconstruct input data for every class in the dataset. In essence, it attempts to reverse-engineer what kind of input data would cause the model to predict a particular class. It achieves this by iteratively optimizing a random input to generate an output that fits the desired target class.

In order to evaluate the effectiveness of the attack, the attack's reconstructed data were compared to the original sample for each class. The Mean Squared Error (MSE) was used as a key metric for the evaluation of the attack. The MSE calculates the average squared difference between the original data and the attack's reconstructed data. Lower MSE values indicate better reconstruction and thus a more successful attack. This measurement metric was computed for every class in which the attack was carried out; therefore, it was computed for both of the dataset's two classification classes. To provide a general evaluation of the attack's efficacy, the average MSE was computed across all classes in addition to the MSE that was generated for each class.

The MSE for every class was computed using Equation (8) as follows:

$$\text{MSE}_c = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

where MSE_c is the MSE for class c, n is the number of features in the sample, y_i is the ith feature of the original sample, and $\hat{y}_i$ is the ith feature of the reconstructed sample.

The average MSE across all classes was calculated using Equation (9) as follows:

$$\text{Average MSE} = \frac{1}{C}\sum_{c=1}^{C}\text{MSE}_c \tag{9}$$

where C is the total number of classes and MSE_c is the MSE for class c.

4.3.4. The Man in the Middle Attack

In this experiment, the attack was executed through a proxy server that intercepted the communication between the clients and the server. In FL, several clients work together to train a common global model. This is when the man in the middle (MITM) attack occurs. Using a proxy server to stand in between the clients and the real server allows for the implementation of the MITM attack. During FL, the participant clients repeatedly request the global model from the actual server. Once the global model has been received from the server, the clients use their local dataset to train it; then, they send the modified, updated weights back to the server to continue the FL process.

The proxy server was a crucial component in the assessment of the applied attack. It performed the attack using a gradient manipulation method. It executed the MITM attack by intercepting the communication from the client providing the server with its updated weights. Specifically, to construct the attacked weights, the proxy server first obtained the original weights from the client by intercepting the communication. Then, it applied the MITM attack by introducing Gaussian noise to the original client's updated weight values, with a standard deviation of 0.1. In this case, two distinct global models were maintained by the actual server: one for the original weights and another for the weights that were attacked. In the case of models that were combined with HE, the proxy directly manipulated the encrypted gradients of HE using the CKKS schema, without decrypting them as HE allows for computations on encrypted data, during the process of training via the addition of encrypted Gaussian noise, enabling the evaluation of the attack in an encrypted environment.

The original and attacked updates (weights/gradients) were both sent to the real server by the proxy server. The server then handled the original and attacked updates independently when receiving them from the proxy server by performing aggregation separately on both the original and attacked updates. It kept track of two global models: the attacked model, which was updated with the aggregated attacked updates, and the original global model, which was updated with the aggregated original updates. After the completion of FL, the server evaluated the effectiveness of the attack by keeping these two distinct global models: the original and attacked global models.

The evaluation was performed by comparing the performance of the final attacked model, which had been impacted by the injected attacked updates, with the performance of the final original global model, which reflected the FL process in the absence of the attack. This enabled the server to measure the impact of the attack by calculating the amount of performance that the attack had caused the model to lose. For both the original global model and the attacked model, the server computed metrics such as the accuracy, precision, recall, and F1-score. Then, it used them to find the percentage drop in each metric measure. The impact of the attack (ASR) was calculated using the percentage decrease for each of the performance metrics of the accuracy, precision, recall, and F1-score using Equation (10).

$$\mathrm{PD} = \frac{(\mathrm{O} - \mathrm{A})}{\mathrm{O}} \times 100 \tag{10}$$

where PD stands for the percentage decrease, O is the original value representing the accuracy, precision, recall, or F1-score metric value before the attack, and A is the attacked value representing the same metric value after the attack.

5. Results and Analysis

5.1. Performance Analysis

Table 4 shows the evaluation of all the models, including the time metric along with the performance metrics of the accuracy, precision, recall, and F1-score. The corresponding classification results for all the models are presented in Table 5. The execution time increased as the number of privacy-preserving methods combinations increased. The FL_only model had the quickest execution time with a minimum value of 82.85 s for the time metric because it only applied FL without any combinations, while the FL_PATE_CKKS_SMPC model had the slowest execution time with a maximum time value of 298.60 s because it included the largest number of combinations among all the models.

Table 4. Overall performance table for all models.

Reduced Dataset	Time (s)	Accuracy	Precision	Recall	F1-Score
Fl_only	82.85	0.9930	0.9930	0.9930	0.9930
FL_SMPC	117.73	0.9950	0.9950	0.9950	0.9950
FL_SMPC_DP	128.54	0.8350	0.8356	0.8350	0.8349
FL_PATE	186.43	0.8530	0.8573	0.8530	0.8526
FL_PATE_SMPC	193.60	0.8650	0.8675	0.8650	0.8648
FL_CKKS	192.87	0.9980	0.9980	0.9980	0.9980
FL_CKKS_DP	204.76	0.7050	0.7196	0.7050	0.7000
FL_CKKS_SMPC	213.08	0.9970	0.9970	0.9970	0.9970
FL_PATE_CKKS	267.34	0.8500	0.8502	0.8500	0.8500
FL_PATE_CKKS_SMPC	298.60	0.8440	0.8442	0.8440	0.8440

Table 5. Overall classification results.

Model (Accuracy)		Precision	Recall	F1-Score
Fl_only (0.99)	malware	0.99	1.00	0.99
	benign	1.00	0.99	0.99
	macro avg	0.99	0.99	0.99
	weighted avg	0.99	0.99	0.99
FL_SMPC (0.99)	malware	0.99	1.00	1.00
	benign	1.00	0.99	0.99
	macro avg	1.00	0.99	0.99
	weighted avg	1.00	0.99	0.99
FL_SMPC_DP (0.83)	malware	0.85	0.81	0.83
	benign	0.82	0.86	0.84
	macro avg	0.84	0.83	0.83
	weighted avg	0.84	0.83	0.83
FL_PATE (0.85)	malware	0.90	0.80	0.84
	benign	0.82	0.91	0.86
	macro avg	0.86	0.85	0.85
	weighted avg	0.86	0.85	0.85
FL_PATE_SMPC (0.86)	malware	0.84	0.91	0.87
	benign	0.90	0.82	0.86
	macro avg	0.87	0.86	0.86
	weighted avg	0.87	0.86	0.86
FL_CKKS (1.00)	malware	1.00	1.00	1.00
	benign	1.00	1.00	1.00
	macro avg	1.00	1.00	1.00
	weighted avg	1.00	1.00	1.00
FL_CKKS_DP (0.70)	malware	0.66	0.83	0.74
	benign	0.78	0.58	0.66
	macro avg	0.72	0.70	0.70
	weighted avg	0.72	0.70	0.70

Table 5. *Cont.*

Model (Accuracy)		Precision	Recall	F1-Score
FL_CKKS_SMPC (1.00)	malware	0.99	1.00	1.00
	benign	1.00	0.99	1.00
	macro avg	1.00	1.00	1.00
	weighted avg	1.00	1.00	1.00
FL_PATE_CKKS (0.85)	malware	0.86	0.84	0.85
	benign	0.84	0.86	0.85
	macro avg	0.85	0.85	0.85
	weighted avg	0.85	0.85	0.85
FL_PATE_CKKS_SMPC (0.84)	malware	0.84	0.86	0.85
	benign	0.85	0.83	0.84
	macro avg	0.84	0.84	0.84
	weighted avg	0.84	0.84	0.84

FL_SMPC was the second-fastest model after FL_only, with a time value of 117.73, because it only combined SMPC and FL with no other additions such as noise. FL_PATE had a moderate time value of 186.43 s as the PATE was executed after the execution of FL, so it increased the training time of the model and resulted in a longer execution time than that of the FL_only model. FL_CKKS also took more time to complete its operations than FL_only due to the added HE operations. FL_CKKS_SMPC exceeded the execution time of FL_CKKS due to the added SMPC operations and secure aggregation using HE and SMPC. FL_CKKS_DP had a higher execution time than FL_CKKS due to noise addition. FL_SMPC_DP took a longer time to complete its operations compared to FL_SMPC because of the addition of regularization and noise.

The FL_CKKS model achieved the highest performance metrics of 99.80% for accuracy, precision, recall, and the F1-score, while FL_CKKS_DP achieved the lowest performance metrics of 70.50% accuracy, 71.96% precision, 70.50% recall, and a 70.00% F1-score. While the PATE enhances privacy, it may decrease the performance due to the noise added to the ensembles of the teachers' votes, so FL_PATE achieved lower performance than FL_only, FL_PATE_CKKS decreased the performance of FL_CKKS, and FL_PATE_CKKS_SMPC reduced the performance of FL_CKKS_SMPC.

FL_SMPC had a performance result that was almost close to that of FL_only because it only applied SMPC during FL with no other data modifications, such as noise addition. FL_SMPC_DP enhanced the privacy of FL_SMPC but decreased its performance due to the addition of noise after the aggregation process. FL_CKKS had performance metrics that were almost close to those of FL_only because it only applied HE operations to FL with no other modifications. On the contrary, FL_CKKS enhanced the performance of FL_only. FL_CKKS_SMPC did not have a significant drop in performance metrics compared to FL_CKKS. On the contrary, FL_CKKS_SMPC achieved results very close to those of FL_CKKS, almost equal to them. FL_CKKS_DP had lower performance compared to FL_CKKS due to the use of DP through noise addition.

5.2. Security Evaluation

Figure 2 visualizes the MSE for all the models in the case of the model inversion attack discussed in Section 4.3.3. As is shown, the MSE was improved through the use of the designed models, proving the importance of applying privacy-preserving techniques along

with FL to protect against model inversion attacks. FL_PATE_SMPC had the best MSE value of 19.267.

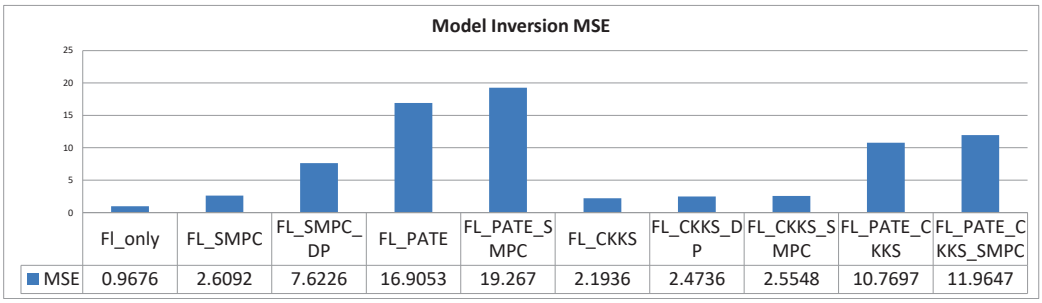

Figure 2. Model inversion MSE.

Figure 3 visualizes the ASR (BASR) of the backdoor attack discussed in Section 4.3.2. As is shown, the ASR decreased when using a combination of privacy-preserving techniques in conjunction with FL. FL_PATE_CKKS_SMPC was the best model for defending against backdoor attacks as it had the smallest backdoor ASR value of 0.0920.

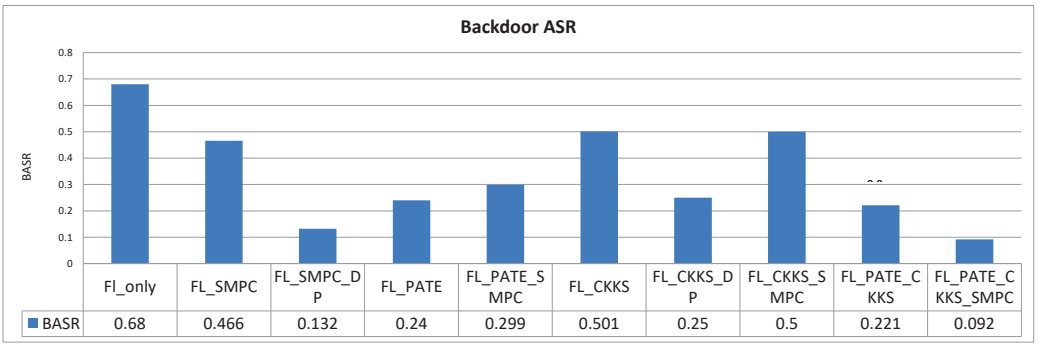

Figure 3. Backdoor ASR.

Figures 4 and 5 visualize the ASR (UPASR and TPASR) for the poisoning attacks discussed in Section 4.3.1. As is shown, the ASR decreased when using a combination of privacy-preserving techniques in conjunction with FL. FL_CKKS_SMPC was the best model for defending against an untargeted poisoning attack as it had the smallest ASR of 0.0010, while FL_CKKS and FL_CKKS_SMPC had the smallest ASR value of 0.0020 for the targeted poisoning attack.

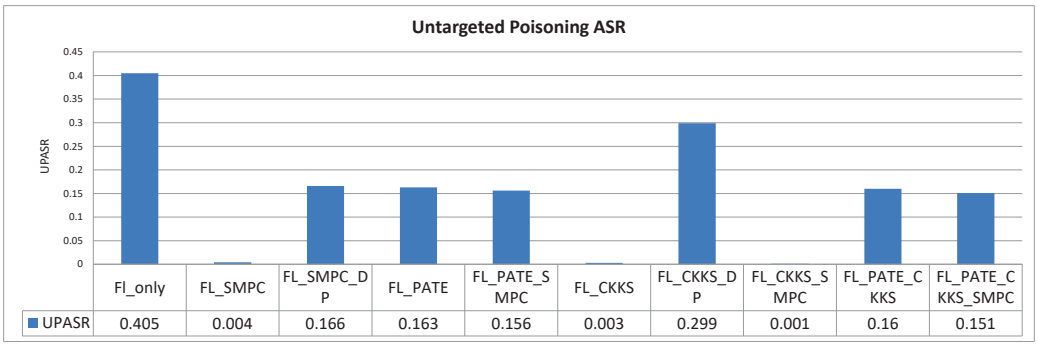

Figure 4. Untargeted poisoning ASR.

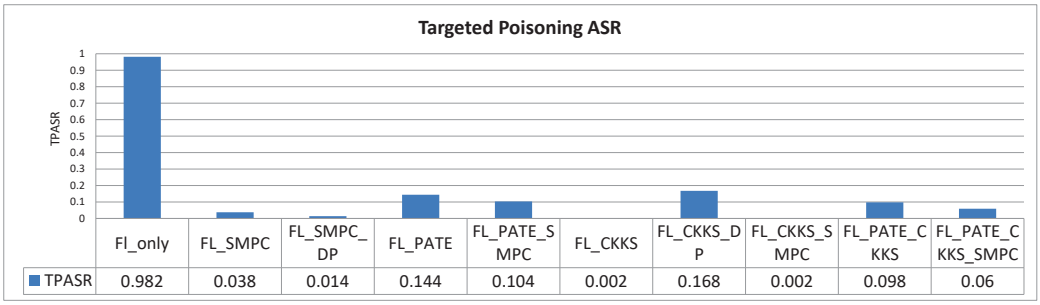

Figure 5. Targeted poisoning ASR.

Using the performance degradation (PD), Table 6 shows the ASR of the MITM attack discussed in Section 4.3.4. As is shown, the performance degradation for the MITM attack decreased with a combination of privacy-preserving techniques together with FL. Based on the reported results, FL_CKKS_PATE_SMPC was the best at resisting the MITM attack, achieving the smallest values of performance degradation for the MITM attack.

Table 6. MITM attack performance degradation (ASR).

Model	Accuracy Degradation	Precision Degradation	Recall Degradation	F1-Score Degradation
Fl_only	48.45%	74.24%	48.45%	65.64%
FL_SMPC	48.51%	74.26%	48.51%	65.67%
FL_SMPC_DP	24.92%	68.52%	24.92%	46.68%
FL_PATE	34.73%	69.01%	34.73%	56.02%
FL_PATE_SMPC	36.79%	68.77%	36.79%	57.77%
FL_CKKS	16.95%	15.80%	16.95%	17.10%
FL_CKKS_DP	25.37%	62.69%	25.37%	50.25%
FL_CKKS_SMPC	11.62%	11.61%	11.62%	11.62%
FL_CKKS_PATE	11.31%	5.24%	11.31%	12.70%
FL_CKKS _PATE_SMPC	1.68%	1.94%	1.68%	1.64%

5.3. Trade-Off Analysis

Figure 6 illustrates a comparison of the training time across all models. The training time increased with the addition of privacy-preserving techniques. Figure 7 visualizes a comparison of the accuracy, precision, recall, and F1-score across all the models. The training time increases and performance may decrease with the addition of privacy-preserving techniques to FL, while this enhances privacy and security by improving the model's robustness against attacks.

Based on the reported results, the integration of privacy preservation techniques and FL provided a stronger privacy guarantee against various attacks. Specifically, for all the evaluated attacks, all the combined models achieved better results compared to the FL_only model. Comparing the FL_PATE_CKKS_SMPC model with the FL_only model, FL_PATE_CKKS_SMPC increased its execution time from 82.85 s for FL_only to 298.60 s. The server and clients of the FL_PATE_CKKS_SMPC model were run 10 times and the average of each step was calculated to obtain the correct percentage. Table 7 shows the time analysis of the FL_PATE_CKKS_SMPC model. This table presents the computational weight of each pipeline step. On the server side, model training took the most time (41.97%), while decryption (3.57%) took the least time. On the client side, communication was the most important factor as it consumed the most time (71.51%), while model loading (0.77%) and

gradient computation (0.25%) used the smallest amount of time. According to the average round times, the server spent significantly less time on each round (4.21%) than the clients (32.85%). The use of the FL_PATE_CKKS_SMPC model decreased the accuracy by approximately 15.01%. The reason behind the decrease in accuracy of the FL_PATE_CKKS_SMPC model was the use of the PATE through noise addition since HE only performed operations on encrypted data and the SMPC used in the FL_PATE_CKKS_SMPC model was the SMPC applied in the FL_SMPC model, which did not use DP. Therefore, HE and SMPC did not participate in the noise reduction of the FL_PATE_CKKS_SMPC model. Specifically, noise introduction through the PATE was the specific factor contributing to the accuracy reduction, not the encryption overhead. In summary, the FL_PATE_CKKS_SMPC model applied the CKKS scheme used in the FL_CKKS model, the SMPC used in the FL_SMPC model, and the PATE used in the FL_PATE model. Both the FL_CKKS and FL_SMPC models did not reduce the accuracy of the base FL_only model, but the FL_PATE model reduced the accuracy of the base FL_only model. Thus, the PATE was the main factor behind the accuracy reduction of the FL_PATE_CKKS_SMPC model. Another explanation is that the FL_PATE_CKKS_SMPC model applied the PATE to the FL_CKKS_SMPC model, which itself did not reduce the accuracy of the base FL_only model, so the PATE was the main factor behind the accuracy reduction of the FL_PATE_CKKS_SMPC model.

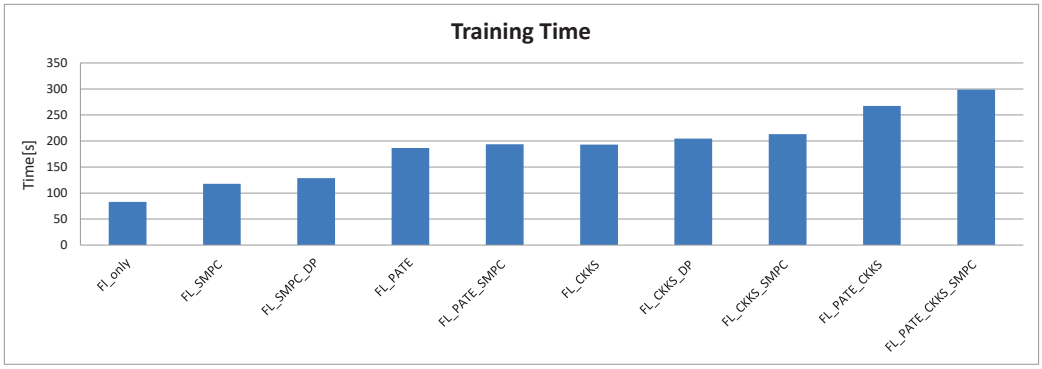

Figure 6. Comparison of models' training times.

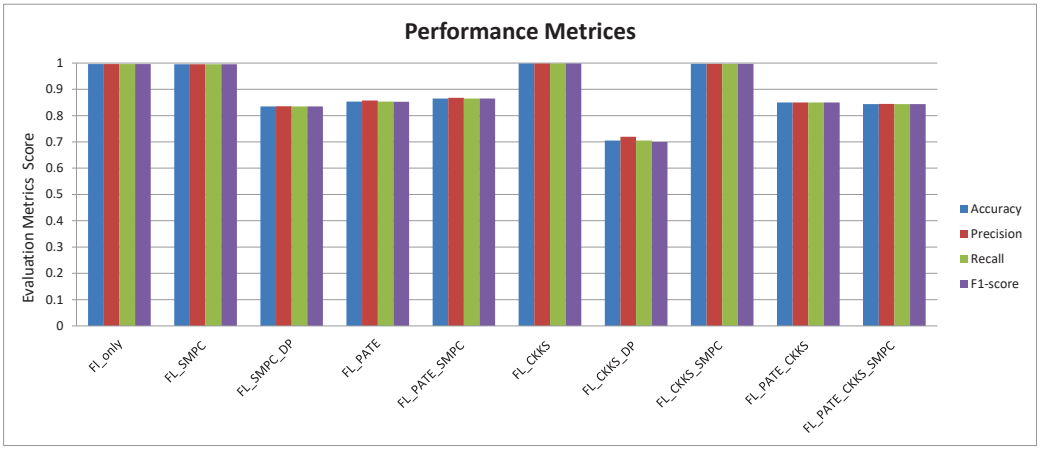

Figure 7. Comparison of models' performance.

Although FL_PATE_CKKS_SMPC increased the execution time and decreased the accuracy of FL_only, it improved the model inversion MSE from 0.9676 for the FL_only model to 11.9647, the backdoor ASR from 0.6800 for FL_only to 0.0920, the untargeted poisoning ASR from 0.4050 for FL_only to 0.151, the targeted poisoning ASR from 0.982 for FL_only to 0.06, and also the MITM attack performance degradation (ASR) from 48.45% accuracy degradation for FL_only to 1.68% for the FL_PATE_CKKS_SMPC model.

Table 7. Time analysis.

Component	Server Side	Client Side
Model Training/PATE Training	41.97% ± 6.51%	13.71% ± 5.14%
Encryption Time	N/A	19.44% ± 8.55%
Decryption Time	3.57% ± 1.22%	N/A
Gradient Calculation Time	N/A	0.25% ± 0.09%
Communication Time	36.48% ± 7.89%	71.51% ± 27.83%
Aggregation Time	7.19% ± 1.99%	N/A
Teacher Model Creation Time	27.13% ± 4.51%	N/A
Model Loading Time	N/A	0.77% ± 0.29%
Average Round Time	4.21% ± 0.77%	32.85% ± 3.13%

6. Discussion

6.1. Insights on Model Execution Time and Privacy Trade-Offs

The results show that the execution time increases as more privacy-preserving techniques are incorporated into the FL model. Notably, the FL_only model, which lacks additional privacy techniques, had the shortest execution time of 82.85 s. As additional privacy-preserving techniques like SMPC, the PATE, and CKKS were integrated, the execution time increased significantly. This was particularly evident in the FL_PATE_CKKS_SMPC model, which had the highest execution time of 298.60 s. This trade-off between privacy and performance underscores the computational cost associated with securing data and enhancing privacy in FL systems. It is crucial to consider this balance when designing privacy-conscious FL systems that must operate efficiently while maintaining data confidentiality.

6.2. Effect of Privacy-Preserving Techniques on Model Accuracy

While privacy-preserving techniques are essential for securing sensitive data, they also introduce noise that impacts the model performance. FL_CKKS, which utilizes homomorphic encryption (HE) without other privacy-enhancing techniques, achieved the highest accuracy and performance metrics, including 99.80% for the accuracy, precision, recall, and F1-score. On the other hand, adding more privacy features such as differential privacy (DP) and Secure Multi-Party Computation (SMPC) resulted in a decline in performance. Specifically, the FL_CKKS_DP model, which includes differential privacy noise, showed a significant drop in metrics, with an accuracy as low as 70.50%. These findings highlight the trade-off between preserving privacy and maintaining the model performance, a critical consideration in fields like healthcare and finance where data security is paramount.

6.3. Privacy and Security Effectiveness Against Attacks

The combination of privacy-preserving techniques notably enhances the robustness of FL models against attacks, such as model inversion and backdoor attacks. For instance, the FL_PATE_SMPC model demonstrated the best defense against model inversion, with the lowest MSE value of 19.267, confirming the effectiveness of the PATE and SMPC in protecting against data leakage. Additionally, the FL_PATE_CKKS_SMPC model achieved the smallest backdoor ASR of 0.0920, further proving the superiority of integrating privacy methods in preventing malicious adversarial activities. This suggests that FL models, when combined with these techniques, can significantly mitigate the risks posed by various privacy and security threats, making them a more viable solution for sensitive applications.

6.4. Impact of Privacy Techniques on Poisoning Attacks

The integration of privacy-preserving techniques also plays a significant role in defending against poisoning attacks, a major concern in FL systems. The results show that FL_CKKS_SMPC was particularly effective in countering untargeted poisoning attacks, achieving the smallest ASR value of 0.0010. Similarly, for targeted poisoning attacks, both FL_CKKS and FL_CKKS_SMPC showed the lowest ASR values of 0.0020. These results highlight the efficacy of combining HE and SMPC in preventing adversaries from corrupting the training process through malicious data poisoning. By incorporating these privacy-enhancing methods, FL systems can ensure the integrity of the model training process, even in the presence of adversarial interference.

6.5. The Impact of Privacy Techniques on the Man in the Middle Attack

The combination of privacy-preserving techniques also improves the resistance of FL models against an MITM attack. The results illustrate that FL_PATE_CKKS_SMPC was the best model to defend against an MITM attack, achieving the lowest performance degradation, with decreases of 1.68% in accuracy, 1.94% in precision, 1.68% in recall, and 1.64% in the F1-score.

6.6. Overall Evaluation and Performance Comparison

The comprehensive performance evaluation of all FL models indicated that while incorporating privacy-preserving techniques can enhance data security, it comes at the cost of an increased computational overhead and slightly reduced performance. However, models like FL_CKKS and FL_CKKS_SMPC strike an optimal balance by achieving high accuracy while maintaining robust privacy protection. The results underscore the importance of carefully selecting privacy techniques that provide adequate protection without severely compromising the model's effectiveness. In the context of real-world applications, especially those handling sensitive data, such as healthcare or financial systems, this trade-off is critical to maintaining both privacy and accuracy.

7. Conclusions and Future Work

This work aimed to provide a novel privacy-enhancing FL framework for malware detection based on a widely used Malware Dataset and an ANN classifier. Without any privacy preservation mechanism, a baseline FL model only achieved performance metrics of 99.30% for the accuracy, precision, recall and F1-score. In our comprehensive analysis, we examined ten different model configurations, testing them against an assortment of machine learning attacks, such as untargeted and targeted poisoning, backdoor injection, model inversion, and MITM attacks. Privacy-preserving mechanisms were integrated into the FL models without any significant performance penalties and increased the robustness of the models against several types of attacks.

From the experiment's results, several top configurations emerged: the FL_PATE_CKKS_SMPC model was the least vulnerable to backdoor attacks (ASR: 0.0920) and to MITM attacks (all metrics' degradation was under 2%); the FL_CKKS_SMPC model showed the highest resistivity to poisoning attacks (ASR: 0.0010 untargeted, 0.0020 targeted) while matching the FL_CKKS model's performance against targeted attacks; and the FL_PATE_SMPC configuration was found to perform the best against model inversion attacks (MSE: 19.267). These results help shed light on the suitability of various privacy-preserving combinations in FL systems, adding to the existing knowledge on secure federated learning systems and offering practical references for privacy-enhancing machine learning applications. The findings suggest a need for continued research to adapt and refine the evaluation framework to encompass newly emerging attack vectors and to develop more robust defense mechanisms.

Future work including an extension to other data types (such as images) or domain-specific datasets (healthcare/financial) would constitute a different research direction.

Author Contributions: Conceptualization, W.K., E.R. and A.S.; Methodology, W.K., E.R. and A.S.; Software, E.S.; Validation, E.S.; Investigation, E.S.; Writing—original draft, E.S.; Writing—review & editing, W.K. and E.R.; Visualization, E.S.; Supervision, W.K., E.R. and A.S.; Project administration, W.K. and A.S.; Funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was funded by the University of Technology and Applied Sciences through the Internal Research Funding Program, grant number IRG-IBRI-25-40.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors are thankful to the anonymous reviewers for their insightful comments and constructive suggestions that greatly improved the quality and clarity of this manuscript. The authors extend their gratitude to Asma M. Alkalbani, University of Technology and Applied Sciences - Ibri, for her invaluable assistance in addressing the proofreading feedback and managing the administrative procedures associated with the funded project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bian, J.; Shen, C.; Xu, J. Federated learning via indirect server-client communications. In Proceedings of the 2023 57th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 22–24 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
2. Zhang, Y.; Zeng, D.; Luo, J.; Fu, X.; Chen, G.; Xu, Z.; King, I. A Survey of Trustworthy Federated Learning: Issues, Solutions, and Challenges. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–47. [CrossRef]
3. Chang, Y.; Zhang, K.; Gong, J.; Qian, H. Privacy-preserving federated learning via functional encryption, revisited. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1855–1869. [CrossRef]
4. Li, Y.; Liu, Z.; Huang, Y.; Xu, P. FedOES: An efficient federated learning approach. In Proceedings of the 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, 24–26 February 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 135–139.
5. Ciceri, O.J.; Astudillo, C.A.; Zhu, Z.; da Fonseca, N.L. Federated learning over next-generation ethernet passive optical networks. *IEEE Netw.* **2022**, *37*, 70–76. [CrossRef]
6. Hussain, G.J.; Manoj, G. Federated learning: A survey of a new approach to machine learning. In Proceedings of the 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichy, India, 16–18 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8.
7. Li, Y.; Xu, G.; Meng, X.; Du, W.; Ren, X. LF3PFL: A Practical Privacy-Preserving Federated Learning Algorithm Based on Local Federalization Scheme. *Entropy* **2024**, *26*, 353. [CrossRef]
8. Sen, J.; Waghela, H.; Rakshit, S. Privacy in Federated Learning. *arXiv* **2024**, arXiv:2408.08904.
9. Li, Y.; Hu, J.; Guo, Z.; Yang, N.; Chen, H.; Yuan, D.; Ding, W. Threats and Defenses in Federated Learning Life Cycle: A Comprehensive Survey and Challenges. *arXiv* **2024**, arXiv:2407.06754.
10. Batool, H.; Anjum, A.; Khan, A.; Izzo, S.; Mazzocca, C.; Jeon, G. A secure and privacy preserved infrastructure for VANETs based on federated learning with local differential privacy. *Inf. Sci.* **2024**, *652*, 119717. [CrossRef]
11. Jin, W.; Yao, Y.; Han, S.; Joe-Wong, C.; Ravi, S.; Avestimehr, S.; He, C. FedML-HE: An efficient homomorphic-encryption-based privacy-preserving federated learning system. *arXiv* **2023**, arXiv:2303.10837.
12. Geng, T.; Liu, J.; Huang, C.T. A Privacy-Preserving Federated Learning Framework for IoT Environment Based on Secure Multi-party Computation. In Proceedings of the 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT), Melbourne, Australia, 24–26 July 2024; pp. 117–122. [CrossRef]
13. Watkins, W.; Wang, H.; Bae, S.; Tseng, H.H.; Cha, J.; Chen, S.Y.C.; Yoo, S. Quantum Privacy Aggregation of Teacher Ensembles (QPATE) for Privacy Preserving Quantum Machine Learning. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 6875–6879.

14. Jiang, Z.; Ni, W.; Zhang, Y. PATE-TripleGAN: Privacy-Preserving Image Synthesis with Gaussian Differential Privacy. *arXiv* **2024**, arXiv:2404.12730.
15. Zhang, Q.; Ma, J.; Lou, J.; Xiong, L.; Jiang, X. Private Semi-supervised Knowledge Transfer for Deep Learning from Noisy Labels. *arXiv* **2022**, arXiv:2211.01628. [CrossRef]
16. Zhao, S.; Zhao, Q.; Zhao, C.; Jiang, H.; Xu, Q. Privacy-enhancing machine learning framework with private aggregation of teacher ensembles. *Int. J. Intell. Syst.* **2022**, *37*, 9904–9920. [CrossRef]
17. Luo, J.; Zhang, Y.; Zhang, J.; Mu, X.; Wang, H.; Yu, Y.; Xu, Z. Secformer: Towards fast and accurate privacy-preserving inference for large language models. *arXiv* **2024**, arXiv:2401.00793.
18. Song, C.; Huang, R.; Hu, S. Private-preserving language model inference based on secure multi-party computation. *Neurocomputing* **2024**, *592*, 127794. [CrossRef]
19. Hosain, M.T.; Abir, M.R.; Rahat, M.Y.; Mridha, M.; Mukta, S.H. Privacy Preserving Machine Learning with Federated Personalized Learning in Artificially Generated Environment. *IEEE Open J. Comput. Soc.* **2024**, *5*, 694–704. [CrossRef]
20. Shen, C.; Zhang, W.; Zhou, T.; Zhang, L. A security-enhanced federated learning scheme based on homomorphic encryption and secret sharing. *Mathematics* **2024**, *12*, 1993. [CrossRef]
21. Liu, X.; Li, H.; Xu, G.; Chen, Z.; Huang, X.; Lu, R. Privacy-enhanced federated learning against poisoning adversaries. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4574–4588. [CrossRef]
22. Xu, G.; Li, H.; Zhang, Y.; Xu, S.; Ning, J.; Deng, R.H. Privacy-Preserving Federated Deep Learning With Irregular Users. *IEEE Trans. Dependable Secur. Comput.* **2022**, *19*, 1364–1381. [CrossRef]
23. Pan, Y.; Ni, J.; Su, Z. FL-PATE: Differentially Private Federated Learning with Knowledge Transfer. In Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM), Kuala Lumpur, Malaysia, 8–12 December 2021; pp. 1–6. [CrossRef]
24. Anastasakis, Z.; Velivassaki, T.H.; Voulkidis, A.; Bourou, S.; Psychogyios, K.; Skias, D.; Zahariadis, T. FREDY: Federated Resilience Enhanced with Differential Privacy. *Future Internet* **2023**, *15*, 296. [CrossRef]
25. Elfares, M.; Reisert, P.; Hu, Z.; Tang, W.; Küsters, R.; Bulling, A. PrivatEyes: Appearance-based Gaze Estimation Using Federated Secure Multi-Party Computation. *Proc. ACM Hum.-Comput. Interact.* **2024**, *8*, 1–23. [CrossRef]
26. Muazu, T.; Mao, Y.; Muhammad, A.U.; Ibrahim, M.; Kumshe, U.M.M.; Samuel, O. A federated learning system with data fusion for healthcare using multi-party computation and additive secret sharing. *Comput. Commun.* **2024**, *216*, 168–182. [CrossRef]
27. Chen, L.; Xiao, D.; Yu, Z.; Zhang, M. Secure and efficient federated learning via novel multi-party computation and compressed sensing. *Inf. Sci.* **2024**, *667*, 120481. [CrossRef]
28. Manh, B.D.; Nguyen, C.H.; Hoang, D.T.; Nguyen, D.N. Homomorphic Encryption-Enabled Federated Learning for Privacy-Preserving Intrusion Detection in Resource-Constrained IoV Networks. *arXiv* **2024**, arXiv:2407.18503.
29. Guo, Y.; Li, L.; Zheng, Z.; Yun, H.; Zhang, R.; Chang, X.; Gao, Z. Efficient and Privacy-Preserving Federated Learning based on Full Homomorphic Encryption. *arXiv* **2024**, arXiv:2403.11519.
30. Gao, Q.; Sun, Y.; Chen, X.; Yang, F.; Wang, Y. An Efficient Multi-Party Secure Aggregation Method Based on Multi-Homomorphic Attributes. *Electronics* **2024**, *13*, 671. [CrossRef]
31. Li, X.; Zhao, H.; Chen, X.; Deng, W. Homomorphic Encryption and Secure Aggregation Based Vertical-Horizontal Federated Learning for Flight Operation Data Sharing. In Proceedings of the 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 29–31 March 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 844–848.
32. Yang, W.; Yang, Y.; Xi, Y.; Zhang, H.; Xiang, W. FLCP: Federated learning framework with communication-efficient and privacy-preserving. *Appl. Intell.* **2024**, *54*, 6816–6835. [CrossRef]
33. Hu, K.; Gong, S.; Zhang, Q.; Seng, C.; Xia, M.; Jiang, S. An overview of implementing security and privacy in federated learning. *Artif. Intell. Rev.* **2024**, *57*, 204. [CrossRef]
34. Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; Gao, Y. A survey on federated learning. *Knowl.-Based Syst.* **2021**, *216*, 106775. [CrossRef]
35. Wen, J.; Zhang, Z.; Lan, Y.; Cui, Z.; Cai, J.; Zhang, W. A survey on federated learning: Challenges and applications. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 513–535. [CrossRef]
36. Chai, D.; Wang, L.; Yang, L.; Zhang, J.; Chen, K.; Yang, Q. A Survey for Federated Learning Evaluations: Goals and Measures. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 5007–5024. [CrossRef]
37. Karras, A.; Karras, C.; Giotopoulos, K.C.; Tsolis, D.; Oikonomou, K.; Sioutas, S. Peer to peer federated learning: Towards decentralized machine learning on edge devices. In Proceedings of the 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Ioannina, Greece, 23–25 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–9.
38. Rachakonda, S.; Moorthy, S.; Jain, A.; Bukharev, A.; Bucur, A.; Manni, F.; Quiterio, T.M.; Joosten, L.; Mendez, N.I. Privacy enhancing and scalable federated learning to accelerate ai implementation in cross-silo and iomt environments. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 744–755. [CrossRef]

39. Qi, P.; Chiaro, D.; Guzzo, A.; Ianni, M.; Fortino, G.; Piccialli, F. Model aggregation techniques in federated learning: A comprehensive survey. *Future Gener. Comput. Syst.* **2024**, *150*, 272–293. [CrossRef]
40. Ryu, M.; Kim, Y.; Kim, K.; Madduri, R.K. APPFL: Open-source software framework for privacy-preserving federated learning. In Proceedings of the 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Lyon, France, 30 May–3 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1074–1083.
41. Pulido-Gaytan, B.; Tchernykh, A.; Tchernykh, A.; Cortés-Mendoza, J.M.; Babenko, M.; Radchenko, G.; Avetisyan, A.; Drozdov, A.Y. Privacy-preserving neural networks with Homomorphic encryption: Challenges and opportunities. *Peer- Netw. Appl.* **2021**, *14*, 1666–1691. [CrossRef]
42. Liu, G.; Furth, N.; Shi, H.; Khreishah, A.; Lee, J.Y.; Ansari, N.; Liu, C.; Jararweh, Y. Federated Learning Aided Deep Convolutional Neural Network Solution for Smart Traffic Management. In Proceedings of the NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium, Miami, FL, USA, 8–12 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
43. Nazir, S.; Kaleem, M. Federated Learning for Medical Image Analysis with Deep Neural Networks. *Diagnostics* **2023**, *13*, 1532–1532. [CrossRef]
44. Gutiérrez, D.; Hassan, H.M.; Landi, L.; Vitaletti, A.; Chatzigiannakis, I. Application of federated learning techniques for arrhythmia classification using 12-lead ECG signals. *arXiv* **2022**, arXiv:2208.10993. [CrossRef]
45. Mothukuri, V.; Khare, P.; Parizi, R.M.; Pouriyeh, S.; Dehghantanha, A.; Srivastava, G. Federated-learning-based anomaly detection for IoT security attacks. *IEEE Internet Things J.* **2021**, *9*, 2545–2554. [CrossRef]
46. Subramanian, N.; Ravi, L.; Shaan, M.J.; Devarajan, M.; Choudhury, T.; Kotecha, K.; Vairavasundaram, S. Securing Mobile Devices from Malware: A Faceoff Between Federated Learning and Deep Learning Models for Android Malware Classification. *J. Comput. Sci.* **2024**, *20*, 254–264. [CrossRef]
47. Panagoda, D.; Malinda, C.; Wijetunga, C.; Rupasinghe, L.; Bandara, B.; Liyanapathirana, C. Application of federated learning in health care sector for malware detection and mitigation using software defined networking approach. In Proceedings of the 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 26–28 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
48. Sikandar, H.S.; Waheed, H.; Tahir, S.; Malik, S.U.; Rafique, W. A detailed survey on federated learning attacks and defenses. *Electronics* **2023**, *12*, 260. [CrossRef]
49. Xia, G.; Chen, J.; Yu, C.; Ma, J. Poisoning attacks in federated learning: A survey. *IEEE Access* **2023**, *11*, 10708–10722. [CrossRef]
50. Liu, P.; Xu, X.; Wang, W. Threats, attacks and defenses to federated learning: Issues, taxonomy and perspectives. *Cybersecurity* **2022**, *5*, 4. [CrossRef]
51. Xu, H.; Shu, T. Defending against model poisoning attack in federated learning: A variance-minimization approach. *J. Inf. Secur. Appl.* **2024**, *82*, 103744. [CrossRef]
52. Nguyen, T.D.; Nguyen, T.; Le Nguyen, P.; Pham, H.H.; Doan, K.D.; Wong, K.S. Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107166. [CrossRef]
53. Chen, Y.; Gui, Y.; Lin, H.; Gan, W.; Wu, Y. Federated learning attacks and defenses: A survey. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 4256–4265.
54. Zhao, J.C.; Bagchi, S.; Avestimehr, S.; Chan, K.S.; Chaterji, S.; Dimitriadis, D.; Li, J.; Li, N.; Nourian, A.; Roth, H.R. Federated Learning Privacy: Attacks, Defenses, Applications, and Policy Landscape-A Survey. *arXiv* **2024**, arXiv:2405.03636.
55. Bradley, M.; Xu, S. A Metric for Machine Learning Vulnerability to Adversarial Examples. In Proceedings of the IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Vancouver, BC, Canada, 10–13 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–2.
56. Yan, H.; Li, X.; Zhang, W.; Wang, R.; Li, H.; Zhao, X.; Li, F.; Lin, X. Automatic evasion of machine learning-based network intrusion detection systems. *IEEE Trans. Dependable Secur. Comput.* **2023**, *21*, 153–167. [CrossRef]
57. Askhatuly, A.; Berdysheva, D.; Yedilkhan, D.; Berdyshev, A. Security Risks of ML Models: Adverserial Machine Learning. In Proceedings of the 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan, 15–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 440–446.
58. Rashid, A.; Such, J. Effectiveness of moving target defenses for adversarial attacks in ml-based malware detection. *arXiv* **2023**, arXiv:2302.00537. [CrossRef]
59. Ikenouchi, H.; Hirose, H.; Uto, T. Backdoor Defense with Colored Patches for Machine Learning Models. In Proceedings of the 2024 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), Okinawa, Japan, 2–5 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
60. Tidjon, L.N.; Khomh, F. Threat assessment in machine learning based systems. *arXiv* **2022**, arXiv:2207.00091.
61. Surekha, M.; Sagar, A.K.; Khemchandani, V. A Comprehensive Analysis of Poisoning Attack and Defence Strategies in Machine Learning Techniques. In Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 9–10 February 2024; IEEE: Piscataway, NJ, USA, 2024, Volume 5; pp. 1662–1668.

62. Chen, L.; Cheng, M.; Huang, H. Backdoor learning on sequence to sequence models. *arXiv* **2023**, arXiv:2305.02424.
63. Dibbo, S.V. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In Proceedings of the 2023 IEEE 36th Computer Security Foundations Symposium (CSF), Dubrovnik, Croatia, 9–13 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 439–456.
64. Zhou, S.; Zhu, T.; Ye, D.; Yu, X.; Zhou, W. Boosting model inversion attacks with adversarial examples. *IEEE Trans. Dependable Secur. Comput.* **2023**, *21*, 1451–1468. [CrossRef]
65. Li, H.; Li, Z.; Wu, S.; Hu, C.; Ye, Y.; Zhang, M.; Feng, D.; Zhang, Y. Seqmia: Sequential-metric based membership inference attack. *arXiv* **2024**, arXiv:2407.15098.
66. Bertran, M.; Tang, S.; Roth, A.; Kearns, M.; Morgenstern, J.H.; Wu, S.Z. Scalable membership inference attacks via quantile regression. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 314–330.
67. Oliynyk, D.; Mayer, R.; Rauber, A. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Comput. Surv.* **2023**, *55*, 1–41. [CrossRef]
68. Rigaki, M.; Garcia, S. Stealing and evading malware classifiers and antivirus at low false positive conditions. *Comput. Secur.* **2023**, *129*, 103192. [CrossRef]
69. Chittibala, D.R.; Jabbireddy, S.R. Security in Machine Learning (ML) Workflows. *Int. J. Comput. Eng.* **2024**, *5*, 52–63. [CrossRef]
70. Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv* **2016**, arXiv:1610.05755.
71. Hannemann, A.; Friedl, B.; Buchmann, E. Differentially Private Multi-Label Learning Is Harder Than You'd Think. In Proceedings of the 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Vienna, Austria, 8–12 July 2024; pp. 40–47. . [CrossRef]
72. Tran, C.; Fioretto, F. On the fairness impacts of private ensembles models. *arXiv* **2023**, arXiv:2305.11807.
73. Malik, J.; Muthalagu, R.; Pawar, P.M. A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls and Technologies. *IEEE Access* **2024**, *12*, 99382–99421. [CrossRef]
74. Mehta, U.; Vekariya, J.; Mehta, M.; Kaur, H.; Kumar, Y. A review of privacy-preserving machine learning algorithms and systems. In *Applied Data Science and Smart Systems*; CRC Press: Boca Raton, FL, USA, 2025; pp. 220–225.
75. Ju, Q.; Xia, R.; Li, S.; Zhang, X. Privacy-preserving classification on deep learning with exponential mechanism. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 39. [CrossRef]
76. Dodwadmath, A.; Stich, S.U. Preserving Privacy with PATE for Heterogeneous Data. In Neural Information Processing Systems Workshop (NeurIPS-W), 2022. Available online: https://publications.cispa.de/articles/conference_contribution/Preserving_privacy_with_PATE_for_heterogeneous_data/24614826?file=43249032 (accessed on 1 February 2025).
77. Hu, H.; Han, Q.; Ma, Z.; Yan, Y.; Xiong, Z.; Jiang, L.; Zhang, Y. PV-PATE: An Improved PATE for Deep Learning with Differential Privacy in Trusted Industrial Data Matrix. In Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Wuhan, China, 6–8 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 477–491.
78. Truong, N.; Sun, K.; Wang, S.; Guitton, F.; Guo, Y. Privacy preservation in federated learning: An insightful survey from the GDPR perspective. *Comput. Secur.* **2021**, *110*, 102402. [CrossRef]
79. Zhou, I.; Tofigh, F.; Piccardi, M.; Abolhasan, M.; Franklin, D.; Lipman, J. Secure Multi-Party Computation for Machine Learning: A Survey. *IEEE Access* **2024**, *12*, 53881–53899. [CrossRef]
80. Khan, T.; Budzys, M.; Nguyen, K.; Michalas, A. Wildest Dreams: Reproducible Research in Privacy-preserving Neural Network Training. *arXiv* **2024**, arXiv:2403.03592. [CrossRef]
81. Parikh, D.; Radadia, S.; Eranna, R.K. Privacy-Preserving Machine Learning Techniques, Challenges And Research Directions. *Int. Res. J. Eng. Technol.* **2024**, *11*, 499.
82. Adelipour, S.; Haeri, M. Private outsourced model predictive control via secure multi-party computation. *Comput. Electr. Eng.* **2024**, *116*, 109208. [CrossRef]
83. Liu, S.; Luo, J.; Zhang, Y.; Wang, H.; Yu, Y.; Xu, Z. Efficient privacy-preserving Gaussian process via secure multi-party computation. *J. Syst. Archit.* **2024**, *151*, 103134. [CrossRef]
84. Tran, A.T.; Luong, T.D.; Pham, X.S. A Novel Privacy-Preserving Federated Learning Model Based on Secure Multi-party Computation. In Proceedings of the International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Kanazawa, Japan, 2–4 November 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 321–333.
85. Krishna, N.; Raju, K.M.; Gowda, V.D.; Arun, G.; Suneetha, S. Homomorphic Encryption and Machine Learning in the Encrypted Domain. In *Innovative Machine Learning Applications for Cryptography*; IGI Global: Hershey, PA, USA, 2024; pp. 173–190.
86. Amorim, I.; Costa, I. Homomorphic Encryption: An Analysis of its Applications in Searchable Encryption. *arXiv* **2023**, arXiv:2306.14407.
87. Gouert, C.; Mouris, D.; Tsoutsos, N. Sok: New insights into fully homomorphic encryption libraries via standardized benchmarks. *Proc. Priv. Enhancing Technol.* **2023**, *2023*, 154–172. [CrossRef]

88. Galymzhankyzy, Z.; Rinatov, I.; Abdiraman, A.; Unaybaev, S. Assessing electoral integrity: Paillier's partial homomorphic encryption in E-voting system. In Proceedings of the 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan, 15–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 194–201.
89. Subramaniyaswamy, V.; Jagadeeswari, V.; Indragandhi, V.; Jhaveri, R.H.; Vijayakumar, V.; Kotecha, K.; Ravi, L. Somewhat homomorphic encryption: Ring learning with error algorithm for faster encryption of IoT sensor signal-based edge devices. *Secur. Commun. Netw.* **2022**, *2022*, 2793998. [CrossRef]
90. van de Haterd, R.; El-Hajj, M. Enhancing Privacy and Security in IoT Environments through Secure Multiparty Computation. In Proceedings of the International Conference on Intelligent Systems and New Applications, Hanoi, Vietnam, 24–25 October 2024; Volume 2; pp. 64–69.
91. Doan, T.V.T.; Messai, M.L.; Gavin, G.; Darmont, J. A survey on implementations of homomorphic encryption schemes. *J. Supercomput.* **2023**, *79*, 15098–15139. [CrossRef]
92. Singh, V.K.; Chauhan, A.S.; Singh, A.; Thakur, R. Homomorphic Encryption: Hands Inside the Gloves. In Proceedings of the 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 21–23 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 248–253.
93. Frimpong, E.; Nguyen, K.; Budzys, M.; Khan, T.; Michalas, A. GuardML: Efficient Privacy-Preserving Machine Learning Services Through Hybrid Homomorphic Encryption. In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, Ávila, Spain, 8–12 April 2024; pp. 953–962.
94. Chillotti, I.; Gama, N.; Georgieva, M.; Izabachène, M. TFHE: Fast fully homomorphic encryption over the torus. *J. Cryptol.* **2020**, *33*, 34–91. [CrossRef]
95. Fan, J.; Vercauteren, F. Somewhat practical fully homomorphic encryption. *Cryptol. Eprint Arch.* **2012** .
96. Cheon, J.H.; Kim, A.; Kim, M.; Song, Y. Homomorphic encryption for arithmetic of approximate numbers. In Proceedings of the Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, 3–7 December 2017; Proceedings, Part I 23; Springer: Berlin/Heidelberg, Germany, 2017; pp. 409–437.
97. Kim, A.; Papadimitriou, A.; Polyakov, Y. Approximate homomorphic encryption with reduced approximation error. In Proceedings of the Cryptographers' Track at the RSA Conference, San Francisco, CA, USA, 1–2 March 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 120–144.
98. Wiryen, Y.B.; Vigny, N.W.A.; Joseph, M.N.; Aimé, F.L. A Comparative Study of BFV and CKKs Schemes to Secure IoT Data Using TenSeal and Pyfhel Homomorphic Encryption Libraries. *Int. J. Smart Secur. Technol. (IJSST)* **2024**, *10*, 1–17. [CrossRef]
99. Patterson, V.L. Hitchhiker's Guide to the TFHE Scheme. *J. Cryptogr. Eng.* **2023**. . [CrossRef]
100. Lee, S.; Lee, G.; Kim, J.W.; Shin, J.; Lee, M.K. HETAL: Efficient privacy-preserving transfer learning with homomorphic encryption. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19010–19035.
101. Zhang, Q.Y.; Wen, Y.W.; Huang, Y.B.; Li, F.P. Secure speech retrieval method using deep hashing and CKKS fully homomorphic encryption. *Multimed. Tools Appl.* **2024**, *83*, 67469–67500. [CrossRef]
102. Reddi, S.; Rao, P.M.; Saraswathi, P.; Jangirala, S.; Das, A.K.; Jamal, S.S.; Park, Y. Privacy-preserving electronic medical record sharing for IoT-enabled healthcare system using fully homomorphic encryption, IOTA, and masked authenticated messaging. *IEEE Trans. Ind. Inform.* **2024**, *20*, 10802–10813. [CrossRef]
103. Kuo, T.H.; Wu, J.L. A High Throughput BFV-Encryption-Based Secure Comparison Protocol. *Mathematics* **2023**, *11*, 1227. [CrossRef]
104. Shen, S.; Yang, H.; Dai, W.; Zhou, L.; Liu, Z.; Zhao, Y. Leveraging GPU in Homomorphic Encryption: Framework Design and Analysis of BFV Variants. *IEEE Trans. Comput.* **2024**, *73*, 2817–2829. [CrossRef]
105. Klemsa, J.; Önen, M.; Akin, Y. A Practical TFHE-Based Multi-Key Homomorphic Encryption with Linear Complexity and Low Noise Growth. In Proceedings of the 28th European Symposium on Research in Computer Security, The Hague, The Netherlands, 25–29 September 2023.
106. Wei, B.; Lu, X.; Wang, R.; Liu, K.; Li, Z.; Wang, K. Thunderbird: Efficient Homomorphic Evaluation of Symmetric Ciphers in 3GPP by combining two modes of TFHE. *IACR Trans. Cryptogr. Hardw. Embed. Syst.* **2024**, *2024*, 530–573. [CrossRef]
107. Rey, V.; Sánchez, P.M.S.; Celdrán, A.H.; Bovet, G. Federated learning for malware detection in IoT devices. *Comput. Netw.* **2022**, *204*, 108693. [CrossRef]
108. Fang, W.; He, J.; Li, W.; Lan, X.; Chen, Y.; Li, T.; Huang, J.; Zhang, L. Comprehensive Android Malware Detection Based on Federated Learning Architecture. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 3977–3990. [CrossRef]
109. Zhang, X.; Wang, C.; Liu, R.; Yang, S. *Federated Rnn-Based Detection of Ransomware Attacks: A Privacy-Preserving Approach*; OSF: Alton, IL, USA, 2024.
110. Jiang, C.; Yin, K.; Xia, C.; Huang, W. Fedhgcdroid: An adaptive multi-dimensional federated learning for privacy-preserving android malware classification. *Entropy* **2022**, *24*, 919. [CrossRef]

111. Nobakht, M.; Javidan, R.; Pourebrahimi, A. SIM-FED: Secure IoT malware detection model with federated learning. *Comput. Electr. Eng.* **2024**, *116*, 109139. [CrossRef]
112. Kalapaaking, A.P.; Stephanie, V.; Khalil, I.; Atiquzzaman, M.; Yi, X.; Almashor, M. Smpc-based federated learning for 6g-enabled internet of medical things. *IEEE Netw.* **2022**, *36*, 182–189. [CrossRef]

information

MDPI

Review

Reliablity and Security for Fog Computing Systems

Egor Shiriaev [1], Tatiana Ermakova [2,*], Ekaterina Bezuglova [1], Maria A. Lapina [3] and Mikhail Babenko [1]

[1] Faculty of Mathematics and Computer Science Named after Prof. N. I. Chervyakov, Department of Computational Mathematics and Cybernetics, North-Caucasus Federal University, 355017 Stavropol, Russia; eshiriaev@ncfu.ru (E.S.); eksbezuglova@ncfu.ru (E.B.); mgbabenko@ncfu.ru (M.B.)
[2] School of Computing, Communication and Business, Hochschule für Technik und Wirtschaft, University of Applied Sciences for Engineering and Economics, 10318 Berlin, Germany
[3] Institute of Digital Development, Department of Information Security of Automated Systems, North-Caucasus Federal University, 355017 Stavropol, Russia; mlapina@ncfu.ru
* Correspondence: tatiana.ermakova@htw-berlin.de

Abstract: Fog computing (FC) is a distributed architecture in which computing resources and services are placed on edge devices closer to data sources. This enables more efficient data processing, shorter latency times, and better performance. Fog computing was shown to be a promising solution for addressing the new computing requirements. However, there are still many challenges to overcome to utilize this new computing paradigm, in particular, reliability and security. Following this need, a systematic literature review was conducted to create a list of requirements. As a result, the following four key requirements were formulated: (1) low latency and response times; (2) scalability and resource management; (3) fault tolerance and redundancy; and (4) privacy and security. Low delay and response can be achieved through edge caching, edge real-time analyses and decision making, and mobile edge computing. Scalability and resource management can be enabled by edge federation, virtualization and containerization, and edge resource discovery and orchestration. Fault tolerance and redundancy can be enabled by backup and recovery mechanisms, data replication strategies, and disaster recovery plans, with a residual number system (RNS) being a promising solution. Data security and data privacy are manifested in strong authentication and authorization mechanisms, access control and authorization management, with fully homomorphic encryption (FHE) and the secret sharing system (SSS) being of particular interest.

Keywords: fog computing; distributed computing systems; reliability; fault tolerance; data security

1. Introduction

Distributed computing systems (DCSs) have evolved over several decades, driven by the need for efficient and scalable computing models [1]. Starting in the 1960s, the issue of distributed computing began to be raised when researchers began to explore the concept of time sharing, allowing multiple users to access the same computer at the same time. This laid the foundation for the development of distributed systems, in which computing resources were shared across multiple nodes.

In the 1970s and 1980s, research efforts focused on the development of local area networks (LANs) and wide area networks (WANs) to connect geographically dispersed computers. This led to the emergence of a client–server architecture, in which a central server handled requests from multiple client devices. DCSs using a client–server architecture have become dominant, enabling resource sharing and centralized data management.

The advent of the Internet in the 1990s led to significant advances in distributed computing. The client–server model has expanded to include web applications, and the World Wide Web (WWW) has become a platform for distributed computing. This era saw the emergence of web services and application programming interfaces (APIs) that enabled interoperability between different systems, making it easier to exchange data and services over the Internet.

Cloud computing (CC) as we know it today emerged in the early 2000s and revolutionized the way computing resources are provided, managed, and used [2]. Cloud technology is based on the idea of providing on-demand access to a pool of computing resources, such as computing power, storage, and software applications delivered over the Internet.

The concept of cloud computing builds on earlier developments in distributed systems, virtualization, and service computing. Virtualization technologies abstract away physical resources, allowing you to create virtual machines (VMs) that can be dynamically provisioned and deallocated as needed.

The key characteristics of cloud computing include on-demand self-service, broad network access, pooling of resources, fast elasticity, and controlled service. Users can access and allocate resources as needed, dynamically scale their usage up or down, and pay for consumed resources on a pay-as-you-go basis [3].

While CC has revolutionized the IT landscape, some applications and use cases have required computing power closer to the network edge. Fog computing (FC) [4] has emerged as an additional cloud computing paradigm to meet these requirements.

FC extends the cloud to the edge of the network, bringing computing, storage, and networking closer to the data source [5]. It uses a distributed architecture consisting of edge devices, gateways, and cloud resources, forming a continuum from the cloud to the edge device.

The FC concept recognizes the limitations of traditional cloud computing in scenarios that require low latency and faster response times, real-time data processing, efficient use of bandwidth, and increased data privacy and security. By processing data closer to the edge, FC reduces the need for extensive data transfer to remote cloud servers, resulting in lower latency and faster response times, as well as the efficient use of bandwidth. Fog devices are also power efficient.

These features make the use of FC attractive in a variety of applications, including IoT [6], smart cities (SCs) [7], automation [8,9], healthcare [10–12], and agriculture [13,14]. In applications such as IoT, SC, and healthcare, a continuous and uninterrupted service is of crucial importance. In the healthcare sector, for example, the failure of a medical monitoring system can have life-threatening consequences. Likewise, in smart cities, the malfunction of traffic management systems could lead to serious disruptions. We call this reliability, which is defined as the continuity of the correct service, e.g., regarding the absence of bugs and the masking of errors [15]. Indeed, shortcomings related to reliability are reported [16]. Furthermore, in healthcare and IoT applications, the protection of personal data is of paramount importance. Here, we refer to the term of security, which covers the ability of the system to keep data confidential, which, in turn, means protecting data from theft, copying, and disclosure of the owner's identity, as well as protection from unauthorized actions [17]. In summary, it can be said that security and reliability are of fundamental importance for the functionality and trustworthiness, and hence wider adoption, of FC systems used in various and critical applications.

The aim of this review was to identify and analyze the requirements for improving the security and reliability of FC, as well as the supporting methods and technologies. For this purpose, we applied the method of a systematic literature research. The literature search was carried out in databases, such as Core, Google Scholar, and Semantic Scholar. The search results were checked for their relevance to the topic, and security and reliability requirements for FC were derived.

The paper is structured as follows: Section 2 discusses the state of the problem. Section 3 presents the background of FC and its application. Section 4 presents the method used, while Section 5 describes the derivation of the requirements. Section 6 provides a discussion of the results and possibilities for future work. Finally, Section 7 summarizes the main conclusions and opportunities for future research.

2. State of the Problem

In work [16], the authors mentioned the disadvantages related to reliability. However, in order to indicate the relevance of the problem under consideration, we considered the following works. In work [18], the authors considered the application of FC in SCs. The main focus of the work was on considering approaches to the integration of FC as the main SC network. The work focused on comparing FC with other types of DCSs, such as CC and edge computing. The work is organized as a review of other research on SC, IoT, and FC topics. The review is general in nature, with a slight refinement for the platform on which FC is deployed. The parameters that were considered in this study were mentioned less often, and the authors suggested paying attention to these parameters, and the use of platform tools is proposed as a solution. A similar trend was also observed in [7]. Here, we considered the integration of smart solutions, their interconnection, and the organization of the infrastructure. The main focus was on device configuration. The issue of fault tolerance was raised from the point of view of the dynamic configuration of device management. However, the authors used standardized data transmission technologies to measure their frequencies, battery life, and messaging speed. Touching on the topic of security, the authors operated with ciphers with a high level of security, but these ciphers were computationally complex. In conclusion, the authors reflected on the need for additional research.

However, the main focus of the authors of work [19] was on the physical model and the construction of a monitoring system, and most of the research was on issues related to the physical embodiment of the system. The research conducted in this study was based on the levels of the OSI model, citing studies at a particular level. In this case, the authors considered a wider range of parameters, including addressing issues of reliability and safety. For example, from the point of view of data protection, the authors characterized FC as difficult to measure. From the point of view of reliability, the authors did not give clear formulations.

Thus, based on several reviews of FC topics, it can be argued that the interest in FC is significant, but at the moment, researchers are focused on building the system as a whole, how to connect devices, which data transfer protocol is more efficient, etc. Although these questions indirectly relate to the subject of this work, a clear answer to the question of "how to organize reliable and secure FC" is still not there. As already mentioned in the previous section, the main purpose of this work was to develop requirements for the reliability and safety of FC. The results will allow other researchers to refer to these requirements to conduct their own FC-related research. Requirements can be in the form of boundary values that will allow you to select the studied methods and algorithms for FC development.

3. Fog Computing and Applications

FC is essentially a distributed computing paradigm that extends cloud capabilities to the edge of the network (Figure 1). While CCs have revolutionized the way computing resources are provided and used, some applications and use cases require low latency, real-time data processing, efficient use of bandwidth, and enhanced privacy and security. FC brings computing, storage, and networking closer to data sources for local processing and analysis [5].

In work [20], the researchers considered the prospects of using FC in real-time applications. Indeed, 8 years later, FC is widely used in smart city networks and the Internet of things. In [21], the authors argued that the use of FC will increase the efficiency of smart city networks based on the Internet of things and proposed a multi-level FC architecture. In [22], the authors also presented the design of a platform with several FC applications.

In [4], the authors conducted research on the applicability of FC for medical purposes, namely, diabetes tracking. The authors compared FC with CC. In their study, they claimed that in terms of diabetes control, FC is more effective than CC in terms of speed, control, and lower network building costs.

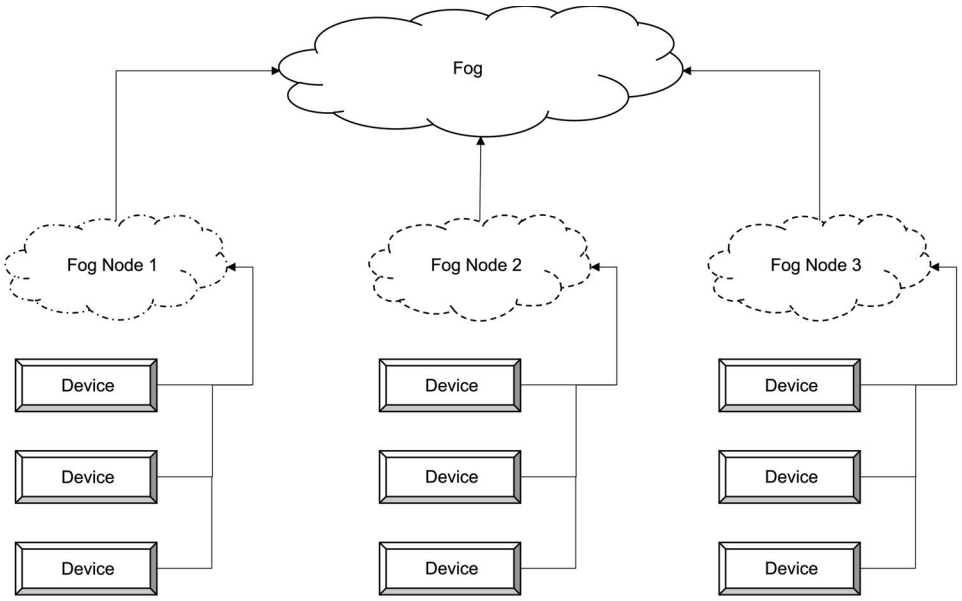

Figure 1. Fog computing.

FC finds use in a variety of applications, including the following:

- IoT: FC has a critical role in IoT deployment [6]. With the proliferation of connected devices generating massive amounts of data, FC brings local processing and analytics to the edge, reducing latency and enabling real-time decision making. This facilitates the efficient filtering and aggregation of data, optimizing the use of network bandwidth and reducing the load on cloud servers.
- Smart cities (SCs): FC plays an important role in the implementation of SC initiatives [7]. When fog nodes (FNs) are deployed throughout the city infrastructure, data from various sources, such as sensors, cameras, and connected devices, can be processed locally. This enables real-time monitoring, analysis, and decision making for applications such as traffic management, waste management, and public security.
- Automation: In industrial environments, FC enables real-time processing and analysis of data for mission-critical applications [8,9]. Edge devices and gateways collect and process data from industrial sensors and equipment, providing local control and monitoring. This reduces latency, provides faster response times, and improves efficiency in industries such as manufacturing, energy, and logistics.
- Healthcare: FC plays a vital role in healthcare systems by facilitating real-time monitoring, analysis, and decision making. Edge devices and gateways can collect and process patient data, enabling timely medical intervention, remote patient monitoring, and personalized healthcare services [10–12]. FC also addresses privacy issues by storing sensitive data locally, ensuring compliance with health regulations.
- Agriculture: In the agricultural sector, FC promotes precision farming and smart farming. Border devices and sensors collect data on soil, weather, and crop conditions to enable local decisions for irrigation, fertilization, and pest control. FC enables real-time analysis and monitoring to optimize resource usage and increase yields [13,14].

Thus, we can say that FC and FSs are a promising directions aimed at automating many areas of human activity. FC and FSs are closely related to the IoT and SC as their components.

4. Method

This study was based on a literature search through databases of scientific publications. Statistics on the number of publications from 2008 (the first mention of FC) to 2024 are listed here, analyzed according to keywords related to reliability and safety (see Figure 2

and Table 1 in the attachment). These include fault tolerance, reliability of functioning, data privacy, data security, and robustness.

Table 1. Scientific papers in keyword databases.

Database	Fault Tolerance	Reliability of Functioning	Data Privacy	Data Security	Robustness
Google Scholar	16,200	16,800	12,700	14,700	14,700
Core	1363	828	792	6014	1740
Semantic Scholar	3050	86	3920	4800	4490

Google Scholar provided the highest number of scientific papers across all keywords, especially in the areas of fault tolerance and functional security. Core and Semantic Scholar also provided a large number of articles, albeit to a lesser extent than Google Scholar (see Figure 2).

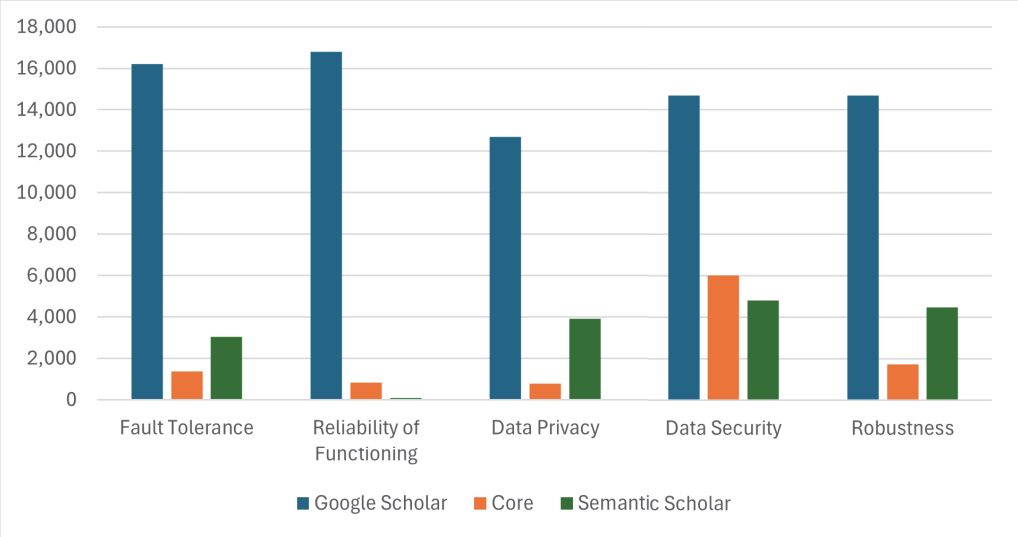

Figure 2. Keyword publication statistics.

As for Google Scholar, the reliability of functioning was the most frequently represented topic with 16,800 contributions, closely followed by fault tolerance with a total of 16,200 contributions. Data security and robustness shared third place with 14,700 publications each. Data privacy was represented by 12,700 publications. According to Core, data security was the most researched topic with 6014 publications. Robustness was in second place with 1740 articles, closely followed by fault tolerance with 1363 publications. The reliability of functioning and data privacy were at the bottom of the rankings with 828 and 792 publications, respectively. In Semantic Scholar, data security was the most researched topic with 4800 papers, closely followed by robustness with 4490 papers, data privacy with 3920 papers, and fault tolerance with 3050 papers. The reliability of functioning, with a count of only 86, completed the ranking (see Figure 3).

The rankings appeared to be inconsistent. The reasons for this, e.g., that the literature databases possibly only contained subsets of the existing publications or had different search mechanisms, can be further analyzed in future work and recommendations derived.

We built our research on the papers with the most citations and downloads (see Tables 2 and 3 below for reliability-related publications and security-related publications analyzed in the manuscript, respectively). They formed the basis for determining and analyzing the requirements for improving the security and reliability of FC.

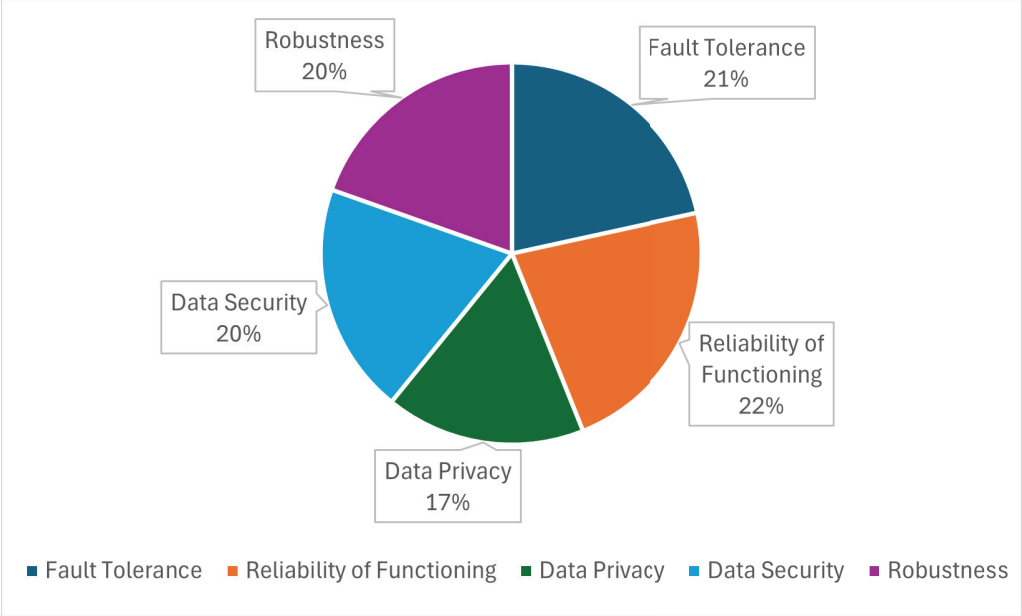

Figure 3. Diagram of the popularity of research topics.

Table 2. Reliability-related publications.

Name	Citations	Downloads	Publisher
Reliability in the utility computing era: Towards reliable fog computing	256	3535	IEEE
Software reliability in the fog computing	24	712	IEEE
A fault-tolerant model for performance optimization of a fog computing system	29	711	IEEE
Distributed fog computing for latency and reliability guaranteed swarm of drones	64	3080	IEEE
A condition of reliability improvement of the system based on the fog-computing concept	10	90	IOPScience
An experimental study of the fog-computing-based systems reliability	7	697	Springer
Fog computing for sustainable smart cities in the IoT era: Caching techniques and enabling technologies—an overview	169	292	Elsevier
Capacity-Aware Edge Caching in Fog Computing Networks	41	1055	IEEE
Fog Computing: An Overview of Big IoT Data Analytics	202	23,455	Hindawi
A Survey on Mobile Edge Computing: The Communication Perspective	4935	63,962	IEEE
A survey on end-edge-cloud orchestrated network computing paradigms: Transparent computing, mobile edge computing, fog computing, and cloudlet	425	6027	ACM
Comparison of Edge Computing Implementations: Fog Computing, Cloudlet and Mobile Edge Computing	609	12,893	IEEE
Fog Computing for Healthcare 4.0 Environment: Opportunities and Challenges	375	477	Elsevier
Fog Computing in Healthcare—A Review and Discussion	360	12,473	IEEE
Fog computing-based IoT for health monitoring system	143	10,020	Hindawi

Table 2. *Cont.*

Name	Citations	Downloads	Publisher
Load-balancing algorithms in cloud computing: A survey	291	345	Elsevier
Game-Theoretic Model for Dynamic Load Balancing in Distributed Systems	63	695	ACM
Analysis of Load Balancing Performance Using Round Robin and IP Hash Algorithm on P4	4	241	IEEE
A Review of Load Balancing in Fog Computing	40	808	IEEE
Hybridization of Firefly and Improved Multi-Objective Particle Swarm Optimization Algorithm for Energy Efficient Load Balancing in Cloud Computing Environments	151	131	Elsevier
Load Balancing in Cloud Computing Environment Using Improved Weighted Round Robin Algorithm for Nonpreemptive Dependent Tasks	237	4767	Hindawi
Fog Computing for Energy-Aware Load Balancing and Scheduling in Smart Factory	330	4976	IEEE
Fog Computing Dynamic Load Balancing Mechanism Based on Graph Repartitioning	213	2294	IEEE
A Blockchain-Based Brokerage Platform for Fog Computing Resource Federation	19	261	IEEE
F-FDN: Federation of Fog Computing Systems for Low Latency Video Streaming	34	534	IEEE
A Review on Container-Based Lightweight Virtualization for Fog Computing	13	1138	IEEE
Feasibility of Fog Computing Deployment Based on Docker Containerization over RaspberryPi	249	2158	ACM
Towards Container Orchestration in Fog Computing Infrastructures	129	2812	IEEE
Foggy: A Platform for Workload Orchestration in a Fog Computing Environment	110	1607	IEEE
Challenges and Solutions in Fog Computing Orchestration	118	2441	IEEE
Topology and Orchestration Specification for Cloud Applications Version 1.0	27	-	OASIS
The Residue Number System	650	1602	ACM
RRNS Base Extension Error-Correcting Code for Performance Optimization of Scalable Reliable Distributed Cloud Data Storage	8	135	IEEE
Parallel Error Correction Algorithm in RNS VLSI Digital Circuits	4	52	IEEE
Correction and Fault Tolerance in RNS-Based Designs. In Residue Number Systems: Theory and Applications	9	979	Springer
A Novel Error Detection and Correction Technique for RNS Based FIR Filters	27	251	IEEE

Table 3. Security-related publications.

Name	Citations	Downloads	Publisher
Fog Computing Security: A Review of Current Applications and Security Solutions	440	49,000	Springer
An Overview of Fog Computing and Its Security Issues	484	-	Wiley
The Fog computing paradigm: Scenarios and security issues	1369	13,222	IEEE
Security and privacy issues of fog computing: A survey	661	7687	Springer
Security and trust issues in Fog computing: A survey	305	335	Elsevier
Centralized fog computing security platform for IoT and cloud in healthcare system	236	-	IGI Global
A Fully Homomorphic Encryption Scheme	3989	-	Stanford
Homomorphic Encryption for Arithmetic of Approximate Numbers	1809	27,000	Springer
Ensemble Method for Privacy-Preserving Logistic Regression Based on Homomorphic Encryption	78	2909	IEEE
Efficient Homomorphic Comparison Methods with Optimal Complexity	116	2914	Springer
FPGA-Based Accelerators of Fully Pipelined Modular Multipliers for Homomorphic Encryption	43	1791	IEEE
Implementation and Performance Evaluation of RNS Variants of the BFV Homomorphic Encryption Scheme	116	1445	IEEE
A Homomorphic Encryption Scheme for Cloud Computing Using Residue Number System	93	879	IEEE
High-Precision Bootstrapping of RNS-CKKS Homomorphic Encryption Using Optimal Minimax Polynomial Approximation and Inverse Sine Function	76	3837	Springer
Secret-Sharing Schemes: A Survey	826	3643	Springer
How to Share a Secret	19,438	29,420	ACM
A Modular Approach to Key Safeguarding	975	1802	IEEE
How to Share a Secret	455	1710	Springer

4.1. Reliability of FC

Reliability can be defined as the continuity of correct service, e.g., in terms of the absence of bugs and the masking of errors [15]. This property can thus refer to both physical characteristics of the network and the software.

In [23], the authors considered the issue of FC reliability based on the reliability of the Internet of things (IoT). Both the IoT and FC assume a system consisting of computing devices with low power consumption.

In [24], the authors carried out a more detailed reliability analysis and divided it into three parts: node reliability, network reliability, and software reliability. The authors proposed methods such as software-defined networking (SDN) and network function virtualization (NFV) as methods to ensure reliability. These methods should improve the reliability of the network if they go through the necessary preliminary stages (network modeling, etc.). To avoid data loss, the authors proposed the use of an adaptive joint protocol based on implicit recognition (AJIA), which is a common reliable and energy-efficient mechanism for packet loss recovery and route quality assessment. In addition, the authors pointed out the need to increase fault tolerance by applying error correction codes.

In [25], the authors present results related to analyzing and predicting the fault tolerance of FC based on continuous Markov chains. Next, the authors present an intelligent FC optimization method, which they call simulated annealing (ISA). The presented method makes it possible to predict which node is most likely to fail in order to take appropriate action.

All the work considered so far assumed the use of the speaker as part of the IoT system. In [26], the authors studied FC in the context of a drone control system and considered the reliability of the system in this domain. In general, the guarantee of reliability is conditioned by the solution of the optimization problem. In this case, the loss of computing nodes (drones) has a higher probability, and thus, the optimization task becomes more complicated. The authors proposed two algorithms: LP-based and proximal Jacobi. For building a foggy drone swarm system, the authors recommended the proximal Jacobi algorithm due to its ratio of problem-solving quality to computational complexity.

The authors of [27] also dealt with the development of FC reliability requirements. In their work, the authors analyzed the dependence of FC reliability depending on the number of computing nodes and workload. The result of the work was calculations that allow us to determine how appropriate it is to use FC on the system under study with specified parameters (the number of nodes and the load on them). The expediency here is the predicted reliability of FC. The authors' contribution to the development of requirements lies in the formulation of which FC can be considered reliable in the theoretical model, namely, such an FC that allows for processing the expected computational load on all nodes without failure.

Ref. [28] took a practical approach and analyzed the reliability of FC depending on the type of system built. In particular, the authors analyzed information systems, such as food chains, healthcare and medical services, mobile-facility-based information systems, smart cities, and UFV monitoring and control. The study consisted of investigating the reliability of the nodes measured at specific points in time. It was carried out with different computational loads: low, medium, and maximum. The authors' research results are useful from the point of view of the practical implementation of FC, namely, for the design of an FC network architecture depending on the expected computational load and the degree and method of distribution of tasks among the computing nodes.

Based on the sources analyzed, the biggest "threat" to FC reliability is the high computational load on the FC nodes. Planning the distribution of the computing load can generally be divided into three requirements:

- Ensuring low latency and response times of the computing nodes;
- Scalability and resource management of a pool of computing nodes;
- Fault tolerance for each individual computing node, as well as general fault tolerance, which can be ensured, among other things, by redundancy of the processed data.

4.2. Security of FC

Security in this study included the ability of the system to keep data confidential, which, in turn, means that data are protected against theft, copying, disclosure of the owner's identity, and unauthorized actions [17].

FC security is based on DCS and CC security since FC is essentially a development of these models. Confidentiality is also important for data in the system, not only in relation to the external environment of the system but also between neighboring nodes. Since the "owners" of computing nodes can be different, an attacker can also gain control over one or more nodes. In addition, FC is a network of low-power nodes, which is a limiting factor on the computational complexity of security methods. Thus, the issue of FC security is relevant and the subject of research in the scientific community.

In [29], the authors provide an analysis of FC security, including CC and edge calculations for the integrity of the study. Given the specifics of the FC system, it can have a wide range of vulnerabilities. Thus, the authors identified 12 categories of security vulnerabilities for FC:

1. Advanced persistent threats (APTs) are cyberattacks that aim to compromise a company's infrastructure in order to steal data and intellectual property.
2. Access control issues (ACIs) can lead to poor management, and any unauthorized user will be able to obtain data and permissions to install software and change configurations.

3. Account hijacking (AH) is when an attack is aimed at hijacking user accounts for malicious purposes. Phishing is a potential method of account hijacking.
4. Denial of service (DoS) is when legitimate users are not allowed to use the system (data and applications) due to excessive use of limited system resources.
5. A data leak (DB) is when an attacker divulges or steals important, protected, or confidential data.
6. Data loss (DL) is the accidental (or malicious) deletion of data from the system. This does not necessarily have to be the result of a cyberattack and may result from a natural disaster.
7. Insecure APIs (IAs): Many cloud service providers provide application programming interfaces (APIs) for use by customers. The security of these APIs is crucial to the security of any implemented applications.
8. System and application vulnerabilities (SAVs) are vulnerabilities that can be exploited as a result of configuration errors in the ad software, which an attacker can use to infiltrate and compromise the system.
9. A malicious insider (MI) is a user who has gained authorized access to the network and system, but intentionally decides to act maliciously.
10. Insufficient due diligence (IDD) often occurs when an organization is in a hurry to adopt, design, and implement a system.
11. Abuse and unfair use (ANU) often occurs when resources are provided free of charge, and attackers use these resources to carry out malicious activities.
12. Problems with shared technologies (STIs) arise from sharing infrastructures, platforms, or applications. For example, the underlying hardware components may not have been designed to provide high-insulation properties.

The authors analyzed how these threats affect various FC-based applications. As a result of their work, the authors cite categories of threats, their types, and possible solutions. In principle, this work can be used as a basis for research in the field of FC security, as well as background information on preparing FC for threat confrontations. However, this work is an overview and does not provide research on the effectiveness of a particular method of protection or methods of attack.

Similar studies are given in the works of [30–34]. In general, these studies can be characterized as follows. The researchers proceeded from the possible applications of FC and focused on their features. On the one hand, this is the right direction since if we compare FC for IoT and FC for healthcare since these are two completely different areas that require different approaches.

However, the very basis of the system is the same for them and the methods of providing the basic level are identical. This conclusion is justified by the conducted research. After analyzing the threats, you can see that, for example, the DB threat for most applications is the same as the ACI threat. Thus, we can say that the main requirement for data security is their confidentiality. Even if an attacker can get hold of one or another part of the data, the data should not have any significance for the attacker.

5. Requirements

Although FC offers many advantages, ensuring reliability and security is crucial for its successful implementation. Based on the review of the literature, we identified the following requirements for the reliability and security of FC.

5.1. Low Delay and Response

FC aims to reduce latency by processing data closer to the edge. It is very important to define acceptable latency thresholds for different applications and use cases. Requirements may vary depending on the need for real-time decision making, data transfer limitations, and application criticality. The following can be considered to optimize the latency and improve the responsiveness in FC environments:

- Edge caching: By caching frequently accessed data and content on edge devices or an FN closest to the end users, latency can be significantly reduced [35]. This allows for faster data and content retrieval as you do not have to traverse the entire network to get to the cloud or remote servers [36]. Edge caching improves the response times for applications such as video streaming, content delivery, and IoT data retrieval.
- Edge analytics: Real-time analytics and decision making at the edge delivers immediate responses without the need to transfer data to a remote server [37]. By deploying simplified analytics and machine learning models directly to the edge, you can instantly obtain insights and actions. Edge analytics is especially useful for time-critical applications, such as industrial automation, smart cities, and healthcare monitoring.
- Mobile edge computing (MEC): MEC brings the power of fog computing to a mobile network infrastructure, delivering low-latency services to mobile devices [38]. By deploying edge computing resources at cell base stations or access points, MEC shortens the distance between mobile devices and computing resources [39,40]. This proximity facilitates real-time applications, such as augmented reality (AR) and virtual reality (VR), where responsiveness is critical. This is useful, for example, in various rehabilitation centers [11,41].

By using these techniques to achieve low latency and a quick response, fog computing environments can perform real-time processing and analysis, reducing the latency and improving the overall system responsiveness. These techniques play an important role in supporting time-critical applications, improving the user experience, and providing a wide range of latency-sensitive services in fog computing deployments.

5.2. Scalability and Resource Management

Fog systems (FSs) must scale easily to meet growing workloads and expanding device connectivity. Determining the requirements for mechanisms for dynamic resource allocation, load balancing and elastic scaling is necessary to ensure the efficient use of resources and system performance. FC systems require robust monitoring and control capabilities to ensure reliability and security. We defined the requirements for real-time monitoring, performance metrics, anomaly detection, and centralized management tools to enable proactive system maintenance, early problem detection, and efficient resource allocation.

- Load balancing: Load-balancing methods evenly distribute computing tasks and network traffic among multiple fog nodes to avoid overloading certain nodes and ensure the optimal use of resources [42]. Load-balancing algorithms consider factors such as the processing capability, network conditions, and node availability to intelligently distribute workloads. This approach avoids bottlenecks, reduces the response time, and improves the system scalability. There are many types of load balancing based on different technologies [43–52].
- Edge federation: Edge federation enables collaboration and resource sharing between multiple fog nodes or edge networks [53]. By forming federated networks, fog computing environments can leverage pooled resources and provide seamless scalability across multiple administrative domains. Federation platforms establish communication protocols, trust mechanisms, and resource sharing agreements to enable dynamic resource allocation and load balancing between federated fog nodes [54].
- Virtualization and containerization: Virtualization technologies, such as hypervisors and VMs, abstract and isolate computing resources, improving the scalability and resource management [55]. Fog nodes can run multiple virtual machines, each with its own sandboxed environment, making efficient use of hardware resources. Containerization using technologies such as Docker [56] provides lightweight and portable environments for applications, enabling efficient deployment, scalability, and resource isolation in fog computing.
- Edge resource discovery and orchestration: Fog computing systems can implement resource discovery mechanisms to determine the available resources in the fog infras-

tructure [57]. These mechanisms facilitate dynamic resource provisioning by allowing fog nodes to efficiently discover and use nearby resources. Orchestration environments coordinate resource provisioning and management, ensuring that workloads are distributed to the appropriate fog nodes based on availability, capability, and proximity [58,59]. There are several effective solutions for orchestration. One of the most interesting is OASIS TOSCA [60], which allows you to effectively manage different containers in distributed systems based on templates, which allows you to combine different approaches. For example, the use of various methods of load balancing.

By leveraging these scaling and resource management techniques, fog computing environments can efficiently handle changing workloads, optimize resource utilization, and ensure the efficient allocation of computing resources. These techniques improve the system scalability, adaptability, and performance, allowing fog computing systems to meet the requirements of various applications and effectively support the dynamic nature of edge computing environments.

5.3. Fault Tolerance and Redundancy

Given the distributed nature of FC, addressing the issues of fault tolerance and redundancy is critical. Defining requirements for error detection and recovery mechanisms and redundancy will help to ensure system reliability and resilience in the face of component failures or network outages. Defining system resiliency and disaster recovery requirements is critical to mitigating potential failures or disasters. Defining backup and recovery mechanisms, data replication strategies, and disaster recovery plans will help to ensure the availability and integrity of data and system functionality.

To meet this requirement, one of the promising areas is the residue number system (RNS), which is a non-positional number system based on modular arithmetic in the residue ring [61]. The RNS has various properties to improve reliability and fault tolerance; in addition, due to natural parallelism and the independence of residuals, the RNS is effectively implemented in distributed systems. The RNS also has self-correcting properties [62–65], which improves the system resiliency. The RNS has many applications, both for reliability and fault tolerance and for security, which is discussed later.

5.4. Data Privacy

FC involves processing sensitive data at the edge, which raises concerns about data privacy and security. Determining the requirements for data encryption, secure communication protocols, access control mechanisms, and compliance with data protection rules is essential to protecting sensitive information. Also, from a privacy perspective, an FS needs strong authentication and authorization mechanisms to ensure that only authorized users and devices can access and interact with the system. Defining requirements for secure user authentication, access control, and privilege management is vital to preventing unauthorized access and protecting system resources.

Maintaining security involves high computing effort and managing the security keys is a more difficult task for CC and FC in terms of data infrastructure compared with centralized systems. Often, centralized security is used for CC. To authenticate the user, the end device communicates with the key exchange server; however, this solution is also ineffective.

One of the promising areas in this field is the so-called fully homomorphic encryption (FHE) [66]. The main feature of FHE is the ability to process data in encrypted form. Many researchers believe that FHE will improve the effectiveness of security and privacy in cloud applications. However, if we consider FC and consider that FNs are low-power devices, FHE is currently not appropriate for FC due to the high computational complexity of some operations [66]. FHE in FC can be examined in Figure 4. However, the current computational complexity is not an unbreakable wall. Many researchers are devoting their work to improving the efficiency of FHE in various applications, such as in works [67–70], where researchers developed an approximate FHE scheme for rational numbers. In addition, the

authors of the papers proposed various modifications of FHE to speed up its operation, for example, by using an RNS [71–73].

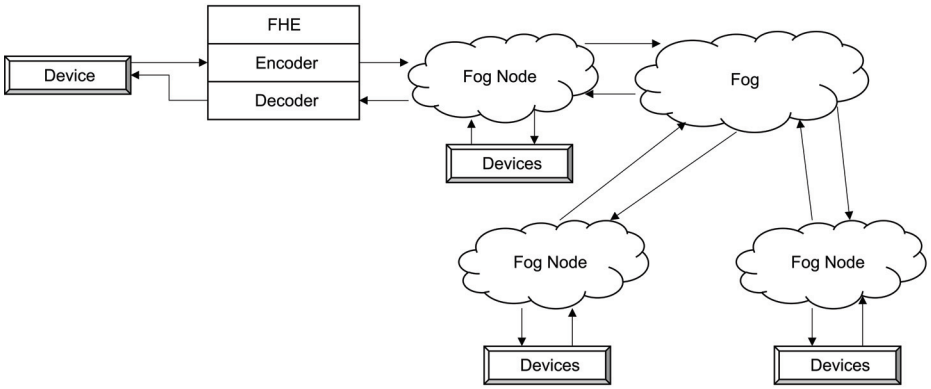

Figure 4. Application of FHE in FC.

Currently, the secret-sharing system (SSS) is the most effective method to fulfill the requirements for security and data protection in FC (Figure 5). The SSS is a security method based on the division of any information into several parts with the possibility of its recovery by a specific group of participants. The SSS is used in conjunction with other security methods to ensure the confidentiality of encryption keys. The SSS is often used in networks with an increased risk of security threats. There are several types of SSS. For example, there are methods that require the capture of all parts of the key (the full SSS) or only a certain proportion (the threshold SSS). Following their emergence, the threshold SSS has almost completely replaced the full SSS.

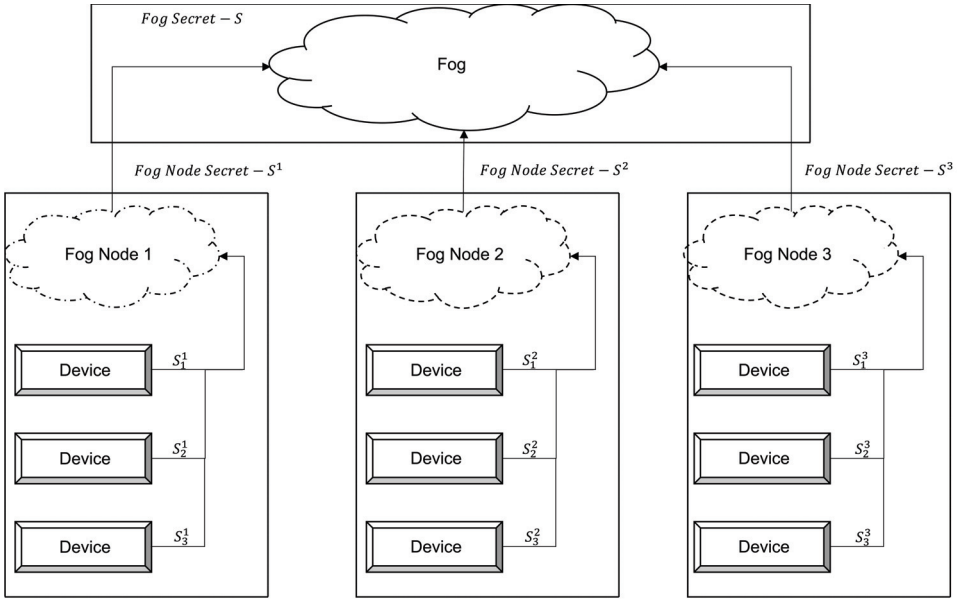

Figure 5. Application of SSS in FC.

Threshold circuits are also of interest for FC. This makes it possible to exchange data together with security keys without the entire system being involved, but only a specific part. This security also fulfills the requirement for FC reliability, as it reduces the load on the network. It is also worth noting that there are SSSs that use an RNS [74,75], which, in turn, allows for a combination of approaches to fulfill both the security and reliability requirements of FC [76–78].

6. Discussion and Future Directions

As a result, the following four key requirements were formulated: (1) low latency and response times; (2) scalability and resource management; (3) fault tolerance and redundancy; and (4) data protection and security. In addition, some methods and technologies are presented that enable FC systems to fulfill these requirements.

Low delay and response can be achieved through edge caching, i.e., caching of frequently accessed data and content on edge devices or the FN closest to the end users [35], edge real-time analyses and decision making at the edge [37], and mobile edge computing [38–40].

Scalability and resource management to avoid overloading certain nodes and ensure optimal resource utilization [42] can be enabled by edge federation, i.e., collaboration and sharing of resources between multiple fog nodes or edge networks [53], virtualization and containerization [55], edge resource discovery and orchestration [57] based on availability, and capabilities of and proximity to the appropriate fog nodes [58,59].

Fault tolerance and redundancy can be enabled by ensuring the availability and integrity of data and system functions with backup and recovery mechanisms, data replication strategies and disaster recovery plans, among other things, with a residual number system (RNS) [61] being a promising solution.

Data security and data privacy are manifested in strong authentication and authorization mechanisms, access control and authorization management, etc., which are ensured by access control mechanisms, secure communication protocols, data encryption, etc., with fully homomorphic encryption (FHE) [66] and a secret sharing system (SSS) being of particular interest.

These requirements can be taken into account when designing FC architectures. Furthermore, on this basis, investigations can be carried out into the extent to which the reliability and security of FC systems and other system properties can be impaired, e.g., to what extent overloads, underloads, and threats can occur. Certain templates can be created based on such data. If more specific templates are available, the developer only needs to specify certain parameters in perspective, e.g., the security level or the load on certain nodes. This will enable the more efficient development and application of FC-based networks and systems.

The focus of future research may include the development of FC architectures with the aim of developing best practices for the use of FC architectures and optimal FC model architectures with respect to the criteria discussed. Simulation tools and mathematical models can be used to evaluate the architectures [79]. For instance, the reliability property can be described by the probability that a system or system component will function successfully up to a certain period of time and can be modeled stochastically, e.g., mean time to failure and probability that a component will fail before reaching time T [15]. Simulators can be developed when mathematical modeling is difficult due to the size, complexity, and heterogeneity of the system [79]. After developing and evaluating these simulation and mathematical models, it is essential to test physical models in the next step in order to validate the results.

Thus, various aspects of FC system security can be studied with a focus on different types of attacks and unauthorized access, as well as the reliability and fault tolerance of FC systems with a focus on the traffic between nodes, queue utilization, and evaluation of device failures. Studies can also focus on the expected computing load with the degree and type of distribution of tasks among the computing nodes [28] and the dependence of FC reliability on the number of computing nodes and the workload, while the expected computing load on all nodes without failure [27] is calculated.

7. Conclusions

In this review, research was conducted on FC reliability and security. This research consisted of an analytical review of the FC literature in order to develop requirements for FC reliability and security. Due to many fundamental differences from CC, not all CC

techniques can be directly transferred to FC. This analysis of the scientific literature on FC showed that the scientific interest in this paradigm is increasing due to the growing interest in smart solutions and local automation. This is due to the fact that FC is an attractive solution for various IoT- and SC-related applications.

Different research groups in the field of reliability have mainly focused on specific FC applications or key system parameters and used them to evaluate the reliability of the system. This study identified key FC parameters that can characterize FC reliability. These parameters have been reformulated to meet the requirements. The result of this part of the study can be considered a list of the formulated requirements themselves, as well as a proposed set of methods and solutions to build an FC that satisfies the requirements.

In the field of FC security, the situation was the opposite; the researchers tried to consider FC security from the point of view of individual threats and possible countermeasures for individual components. Some of the threats analyzed in this paper fall under the category of reliability and fault tolerance. On this basis, a single requirement, "data confidentiality", was defined in the field of security. Since in the case of FC, data confidentiality is guaranteed, it will be difficult for an attacker to successfully attack the system to steal important data. This paper also presents methods to ensure the required level of confidentiality.

In the future, practical research is planned, namely, the construction of FC architectures based on the results of the conducted research in order to work out FC system templates. Also, there is planned experimental study of the developed architectures on the subject of conformity to the developed requirements of reliability and security, interchangeability of architectures and templates, and methods of reliability and security.

Author Contributions: Conceptualization, M.B. and T.E.; methodology, E.S. and E.B.; validation, E.S. and M.A.L.; formal analysis, T.E. and E.B.; investigation, M.B.; resources, T.E.; writing—original draft preparation, E.S.; writing—review and editing, T.E.; visualization, E.S. and E.B.; supervision, T.E.; project administration, M.B.; funding acquisition, M.B. All authors read and agreed to the published version of this manuscript.

Funding: This research was supported by the Russian Science Foundation Grant No. 22-71-10046, 414 (https://rscf.ru/en/project/22-71-10046/ (accessed on 24 April 2024)).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AG	Augmented reality
DCS	Distributed computing system
CC	Cloud computing
FC	Fog computing
FHE	Fully homomorphic encryption
FN	Fog node
FS	Fog system
IaaS	Infrastructure as a service
IoT	Internet of things
MEC	Mobile edge computing
PaaS	Platform as a service
RNS	Residue number system
SaaS	Software as a service
SC	Smart city
SSS	Secret sharing system

VR Virtual reality

References

1. Singh, K.; Alam, M.; Kumar, S. A Survey of Static Scheduling Algorithm for Distributed Computing System. *Int. J. Comput. Appl.* **2015**, *129*, 25–30. [CrossRef]
2. Kratzke, N. A Brief History of Cloud Application Architectures. *Appl. Sci.* **2018**, *8*, 1368. [CrossRef]
3. Mell, P.; Grance, T. The NIST Definition of Cloud Computing. National Institute of Standards and Technology. 2011. Available online: https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf (accessed on 22 April 2024).
4. Klonoff, D.C. Fog Computing and Edge Computing Architectures for Processing Data From Diabetes Devices Connected to the Medical Internet of Things. *J. Diabetes Sci. Technol.* **2017**, *11*, 647–652. [CrossRef]
5. Abouaomar, A.; Cherkaoui, S.; Kobbane, A.; Dambri, O.A. A Resources Representation for Resource Allocation in Fog Computing Networks. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
6. Sarkar, S.; Chatterjee, S.; Misra, S. Assessment of the Suitability of Fog Computing in the Context of Internet of Things. *IEEE Trans. Cloud Comput.* **2015**, *6*, 46–59. [CrossRef]
7. Perera, C.; Qin, Y.; Estrella, J.C.; Reiff-Marganiec, S.; Vasilakos, A.V. Fog Computing for Sustainable Smart Cities: A Survey. *ACM Comput. Surv.* **2017**, *50*, 32:1–32:43. [CrossRef]
8. Pop, P.; Raagaard, M.A.L.; Gutierrez, M.; Steiner, W. Enabling Fog Computing for Industrial Automation Through Time-Sensitive Networking (TSN). *IEEE Commun. Stand. Mag.* **2018**, *2*, 55–61. [CrossRef]
9. Rani, S.; Kataria, A.; Chauhan, M. Fog Computing in Industry 4.0: Applications and Challenges—A Research Roadmap. In *Energy Conservation Solutions for Fog-Edge Computing Paradigms*; Tiwari, R., Mittal, M., Goyal, L.M., Eds.; Springer: Singapore, 2022; pp. 173–190, ISBN 9789811634482.
10. Kraemer, F.A.; Braten, A.E.; Tamkittikhun, N.; Palma, D. Fog Computing in Healthcare—A Review and Discussion. *IEEE Access* **2017**, *5*, 9206–9222. [CrossRef]
11. Kumari, A.; Tanwar, S.; Tyagi, S.; Kumar, N. Fog Computing for Healthcare 4.0 Environment: Opportunities and Challenges. *Comput. Electr. Eng.* **2018**, *72*, 1–13. [CrossRef]
12. Shi, Y.; Ding, G.; Wang, H.; Roman, H.E.; Lu, S. The Fog Computing Service for Healthcare. In Proceedings of the 2015 2nd International Symposium on Future Information and Communication Technologies for Ubiquitous HealthCare (Ubi-HealthTech), Beijing, China, 28–30 May 2015; pp. 1–5.
13. Guardo, E.; Stefano, A.D.; Corte, A.L.; Sapienza, M.; Scatà, M. A Fog Computing-Based IoT Framework for Precision Agriculture. *J. Internet Technol.* **2018**, *19*, 1401–1411.
14. Hsu, T.-C.; Yang, H.; Chung, Y.-C.; Hsu, C.-H. A Creative IoT Agriculture Platform for Cloud Fog Computing. *Sustain. Comput. Inform. Syst.* **2020**, *28*, 100285. [CrossRef]
15. Berger, C.; Eichhammer, P.; Reiser, H.P.; Domaschka, J.; Hauck, F.J.; Habiger, G. A Survey on Resilience in the IoT: Taxonomy, Classification, and Discussion of Resilience Mechanisms. *ACM Comput. Surv.* **2022**, *54*, 1–39. [CrossRef]
16. Yi, S.; Li, C.; Li, Q. A Survey of Fog Computing: Concepts, Applications and Issues. In Proceedings of the 2015 Workshop on Mobile Big Data, Hangzhou, China, 21 June 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 37–42.
17. Soomro, Z.A.; Shah, M.H.; Ahmed, J. Information Security Management Needs More Holistic Approach: A Literature Review. *Int. J. Inf. Manag.* **2016**, *36*, 215–225. [CrossRef]
18. Songhorabadi, M.; Rahimi, M.; MoghadamFarid, A.; Kashani, M.H. Fog Computing Approaches in IoT-Enabled Smart Cities. *J. Netw. Comput. Appl.* **2023**, *211*, 103557. [CrossRef]
19. Sabireen, H.; Neelanarayanan, V. A Review on Fog Computing: Architecture, Fog with IoT, Algorithms and Research Challenges. *ICT Express* **2021**, *7*, 162–176.
20. Peter, N. Fog computing and its real time applications. *Int. J. Emerg. Technol. Adv. Eng.* **2015**, *5* , 266–269.
21. Zhang, C. Design and Application of Fog Computing and Internet of Things Service Platform for Smart City. *Future Gener. Comput. Syst.* **2020**, *112*, 630–640. [CrossRef]
22. Yi, S.; Hao, Z.; Qin, Z.; Li, Q. Fog Computing: Platform and Applications. In Proceedings of the 2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb), Washington, DC, USA, 12–13 November 2015; pp. 73–78.
23. Madsen, H.; Burtschy, B.; Albeanu, G.; Popentiu-Vladicescu, F.L. Reliability in the Utility Computing Era: Towards Reliable Fog Computing. In Proceedings of the 2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP), Bucharest, Romania, 7–9 July 2013; pp. 43–46.
24. Popentiu-Vladicescu, F.; Albeanu, G. Software Reliability in the Fog Computing. In Proceedings of the 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), Karachi, Pakistan, 5–7 April 2017; pp. 1–4.
25. Zhang, P.; Chen, Y.; Zhou, M.; Xu, G.; Huang, W.; Al-Turki, Y.; Abusorrah, A. A Fault-Tolerant Model for Performance Optimization of a Fog Computing System. *IEEE Internet Things J.* **2022**, *9*, 1725–1736. [CrossRef]
26. Hou, X.; Ren, Z.; Wang, J.; Zheng, S.; Cheng, W.; Zhang, H. Distributed Fog Computing for Latency and Reliability Guaranteed Swarm of Drones. *IEEE Access* **2020**, *8*, 7117–7130. [CrossRef]

27. Melnik, E.V.; Klimenko, A.B. A Condition of Reliability Improvement of the System Based on the Fog-Computing Concept. *J. Phys. Conf. Ser.* **2020**, *1661*, 012007. [CrossRef]
28. Klimenko, A.B.; Melnik, E.V. An Experimental Study of the Fog-Computing-Based Systems Reliability. In *Artificial Intelligence and Bioinspired Computational Methods*; Silhavy, R., Ed.; Springer International Publishing: Cham, Switzerland, 2020; pp. 438–449.
29. Khan, S.; Parkinson, S.; Qin, Y. Fog Computing Security: A Review of Current Applications and Security Solutions. *J. Cloud Comp.* **2017**, *6*, 19. [CrossRef]
30. Stojmenovic, I.; Wen, S.; Huang, X.; Luan, H. An Overview of Fog Computing and Its Security Issues. *Concurr. Comput. Pract. Exper.* **2016**, *28*, 2991–3005. [CrossRef]
31. Stojmenovic, I.; Wen, S. The Fog Computing Paradigm: Scenarios and Security Issues. In Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, 7–10 September 2014; pp. 1–8.
32. Yi, S.; Qin, Z.; Li, Q. Security and Privacy Issues of Fog Computing: A Survey. In *Wireless Algorithms, Systems, and Applications*; Xu, K., Zhu, H., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 685–695.
33. Zhang, P.; Zhou, M.; Fortino, G. Security and Trust Issues in Fog Computing: A Survey. *Future Gener. Comput. Syst.* **2018**, *88*, 16–27. [CrossRef]
34. Thota, C.; Sundarasekar, R.; Manogaran, G.; Varatharajan, R.; Priyan, M.K. Centralized Fog Computing Security Platform for IoT and Cloud in Healthcare System. In *Fog Computing: Breakthroughs in Research and Practice*; IGI Global: Pennsylvania, PA, USA, 2018; pp. 365–378, ISBN 978-1-5225-5649-7.
35. Zahmatkesh, H.; Al-Turjman, F. Fog Computing for Sustainable Smart Cities in the IoT Era: Caching Techniques and Enabling Technologies—An Overview. *Sustain. Cities Soc.* **2020**, *59*, 102139. [CrossRef]
36. Li, Q.; Zhang, Y.; Li, Y.; Xiao, Y.; Ge, X. Capacity-Aware Edge Caching in Fog Computing Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 9244–9248. [CrossRef]
37. Anawar, M.R.; Wang, S.; Azam Zia, M.; Jadoon, A.K.; Akram, U.; Raza, S. Fog Computing: An Overview of Big IoT Data Analytics. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 7157192. [CrossRef]
38. Mao, Y.; You, C.; Zhang, J.; Huang, K.; Letaief, K.B. A Survey on Mobile Edge Computing: The Communication Perspective. *IEEE Commun. Surv. Tutorials* **2017**, *19*, 2322–2358. [CrossRef]
39. Ren, J.; Zhang, D.; He, S.; Zhang, Y.; Li, T. A Survey on End-Edge-Cloud Orchestrated Network Computing Paradigms: Transparent Computing, Mobile Edge Computing, Fog Computing, and Cloudlet. *ACM Comput. Surv.* **2019**, *52*, 125:1–125:36. [CrossRef]
40. Dolui, K.; Datta, S.K. Comparison of Edge Computing Implementations: Fog Computing, Cloudlet and Mobile Edge Computing. In Proceedings of the 2017 Global Internet of Things Summit (GIoTS), Geneva, Switzerland, 6–9 June 2017; pp. 1–6.
41. Paul, A.; Pinjari, H.; Hong, W.-H.; Seo, H.C.; Rho, S. Fog Computing-Based IoT for Health Monitoring System. *J. Sens.* **2018**, *2018*, e1386470. [CrossRef]
42. Jafarnejad Ghomi, E.; Masoud Rahmani, A.; Nasih Qader, N. Load-Balancing Algorithms in Cloud Computing: A Survey. *J. Netw. Comput. Appl.* **2017**, *88*, 50–71. [CrossRef]
43. Aote, S.S.; Kharat, M.U. A Game-Theoretic Model for Dynamic Load Balancing in Distributed Systems. In Proceedings of the International Conference on Advances in Computing, Communication and Control, Bangalore, India, 28–29 December 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 235–238.
44. Baihaqi, M.R.; Negara, R.M.; Tulloh, R. Analysis of Load Balancing Performance Using Round Robin and IP Hash Algorithm on P4. In Proceedings of the 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 8 December 2022; pp. 93–98.
45. Mishra, S.K.; Sahoo, B.; Parida, P.P. Load Balancing in Cloud Computing: A Big Picture. *J. King Saud Univ. Comput. Inf. Sci.* **2020**, *32*, 149–158. [CrossRef]
46. Kumar, P.; Kumar, R. Issues and Challenges of Load Balancing Techniques in Cloud Computing: A Survey. *ACM Comput. Surv.* **2019**, *51*, 120:1–120:35. [CrossRef]
47. Jader, O.; Zeebaree, S.; Zebari, R. A State Of Art Survey For Web Server Performance Measurement And Load Balancing Mechanisms. *Int. J. Sci. Technol. Res.* **2019**, *8*, 535–543.
48. Chandak, A.; Ray, N.K. A Review of Load Balancing in Fog Computing. In Proceedings of the 2019 International Conference on Information Technology (ICIT), Shanghai, China, 20–23 December 2019; pp. 460–465.
49. Devaraj, A.F.S.; Elhoseny, M.; Dhanasekaran, S.; Lydia, E.L.; Shankar, K. Hybridization of Firefly and Improved Multi-Objective Particle Swarm Optimization Algorithm for Energy Efficient Load Balancing in Cloud Computing Environments. *J. Parallel Distrib. Comput.* **2020**, *142*, 36–45. [CrossRef]
50. Devi, D.C.; Uthariaraj, V.R. Load Balancing in Cloud Computing Environment Using Improved Weighted Round Robin Algorithm for Nonpreemptive Dependent Tasks. *Sci. World J.* **2016**, *2016*, e3896065. [CrossRef]
51. Wan, J.; Chen, B.; Wang, S.; Xia, M.; Li, D.; Liu, C. Fog Computing for Energy-Aware Load Balancing and Scheduling in Smart Factory. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4548–4556. [CrossRef]
52. Ningning, S.; Chao, G.; Xingshuo, A.; Qiang, Z. Fog Computing Dynamic Load Balancing Mechanism Based on Graph Repartitioning. *China Commun.* **2016**, *13*, 156–164. [CrossRef]

53. Savi, M.; Santoro, D.; Di Meo, K.; Pizzolli, D.; Pincheira, M.; Giaffreda, R.; Cretti, S.; Kum, S.; Siracusa, D. A Blockchain-Based Brokerage Platform for Fog Computing Resource Federation. In Proceedings of the 2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), Paris, France, 24–27 February 2020; pp. 147–149.
54. Veillon, V.; Denninnart, C.; Salehi, M.A. F-FDN: Federation of Fog Computing Systems for Low Latency Video Streaming. In Proceedings of the 2019 IEEE 3rd International Conference on Fog and Edge Computing (ICFEC), Larnaca, Cyprus, 14–17 May 2019; pp. 1–9.
55. Sri Raghavendra, M.; Chawla, P. A Review on Container-Based Lightweight Virtualization for Fog Computing. In Proceedings of the 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 29–31 August 2018; pp. 378–384.
56. Bellavista, P.; Zanni, A. Feasibility of Fog Computing Deployment Based on Docker Containerization over RaspberryPi. In Proceedings of the 18th International Conference on Distributed Computing and Networking, Hyderabad, India, 5–7 January 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–10.
57. Hoque, S.; De Brito, M.S.; Willner, A.; Keil, O.; Magedanz, T. Towards Container Orchestration in Fog Computing Infrastructures. In Proceedings of the 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, Italy, 4–8 July 2017; Volume 2, pp. 294–299.
58. Santoro, D.; Zozin, D.; Pizzolli, D.; De Pellegrini, F.; Cretti, S. Foggy: A Platform for Workload Orchestration in a Fog Computing Environment. In Proceedings of the 2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Hong Kong, 11–14 December 2017; pp. 231–234.
59. Jiang, Y.; Huang, Z.; Tsang, D.H.K. Challenges and Solutions in Fog Computing Orchestration. *IEEE Netw.* **2018**, *32*, 122–129. [CrossRef]
60. OASIS Topology and Orchestration Specification for Cloud Applications (TOSCA) TC. Topology and Orchestration Specification for Cloud Applications Version 1.0. 2013. Available online: http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html (accessed on 22 April 2024).
61. Valueva, M.V.; Nagornov, N.N.; Lyakhov, P.A.; Valuev, G.V.; Chervyakov, N.I. Application of the Residue Number System to Reduce Hardware Costs of the Convolutional Neural Network Implementation. *Math. Comput. Simul.* **2020**, *177*, 232–243. [CrossRef]
62. Babenko, M.; Tchernykh, A.; Pulido-Gaytan, B.; Cortés-Mendoza, J.M.; Shiryaev, E.; Golimblevskaia, E.; Avetisyan, A.; Nesmachnow, S. RRNS Base Extension Error-Correcting Code for Performance Optimization of Scalable Reliable Distributed Cloud Data Storage. In Proceedings of the 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Portland, OR, USA, 17–21 June 2021; pp. 548–553.
63. Tay, T.F.; Chang, C.-H. A New Algorithm for Single Residue Digit Error Correction in Redundant Residue Number System. In Proceedings of the 2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne, Australia, 1–5 June 2014; pp. 1748–1751. [CrossRef]
64. Ananda Mohan, P.V. Error Detection, Correction and Fault Tolerance in RNS-Based Designs. In *Residue Number Systems: Theory and Applications*; Mohan, P.V.A., Ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 163–175, ISBN 978-3-319-41385-3.
65. Chang, C.-H.; Molahosseini, A.S.; Zarandi, A.A.E.; Tay, T.F. Residue Number Systems: A New Paradigm to Datapath Optimization for Low-Power and High-Performance Digital Signal Processing Applications. *IEEE Circ. Syst. Mag.* **2015**, *15*, 26–44. [CrossRef]
66. Gentry, C. A Fully Homomorphic Encryption Scheme. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 2009.
67. Cheon, J.H.; Kim, A.; Kim, M.; Song, Y. Homomorphic Encryption for Arithmetic of Approximate Numbers. In *Advances in Cryptology—ASIACRYPT 2017*; Takagi, T., Peyrin, T., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 409–437.
68. Cheon, J.H.; Kim, D.; Kim, Y.; Song, Y. Ensemble Method for Privacy-Preserving Logistic Regression Based on Homomorphic Encryption. *IEEE Access* **2018**, *6*, 46938–46948. [CrossRef]
69. Cheon, J.H.; Kim, D.; Kim, D. Efficient Homomorphic Comparison Methods with Optimal Complexity. In *Advances in Cryptology—ASIACRYPT 2020*; Moriai, S., Wang, H., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 221–256.
70. Kim, S.; Lee, K.; Cho, W.; Cheon, J.H.; Rutenbar, R.A. FPGA-Based Accelerators of Fully Pipelined Modular Multipliers for Homomorphic Encryption. In Proceedings of the 2019 International Conference on ReConFigurable Computing and FPGAs (ReConFig), Cancun, Mexico, 9–11 December 2019; pp. 1–8.
71. Al Badawi, A.; Polyakov, Y.; Aung, K.M.M.; Veeravalli, B.; Rohloff, K. Implementation and Performance Evaluation of RNS Variants of the BFV Homomorphic Encryption Scheme. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 941–956. [CrossRef]
72. Gomathisankaran, M.; Tyagi, A.; Namuduri, K. HORNS: A Homomorphic Encryption Scheme for Cloud Computing Using Residue Number System. In Proceedings of the 2011 45th Annual Conference on Information Sciences and Systems, Baltimore, MD, USA, 23–25 March 2011; pp. 1–5.
73. Lee, J.-W.; Lee, E.; Lee, Y.; Kim, Y.-S.; No, J.-S. High-Precision Bootstrapping of RNS-CKKS Homomorphic Encryption Using Optimal Minimax Polynomial Approximation and Inverse Sine Function. In *Advances in Cryptology—EUROCRYPT 2021*; Canteaut, A., Standaert, F.-X., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 618–647.

74. Goyal, V.; Kumar, A. Non-Malleable Secret Sharing. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, Los Angeles, CA, USA, 25–29 June 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 685–698.
75. Applebaum, B.; Beimel, A.; Farràs, O.; Nir, O.; Peter, N. Secret-Sharing Schemes for General and Uniform Access Structures. In *Advances in Cryptology—EUROCRYPT 2019*; Ishai, Y., Rijmen, V., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 441–471.
76. Gladkov, A.; Gladkova, N.; Kucherov, N. Analytical Review of Methods for Detection, Localization and Error Correction in the Residue Number System. In *Mathematics and its Applications in New Computer Systems*; Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 507–514.
77. Gladkov, A.; Shiriaev, E.; Tchernykh, A.; Deryabin, M.; Bezuglova, E.; Valuev, G.; Babenko, M. SNS-Based Secret Sharing Scheme for Security of Smart City Communication Systems. In *Smart Cities*; Nesmachnow, S., Hernández Callejo, L., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 248–263.
78. Boyle, E.; Gilboa, N.; Ishai, Y. Function Secret Sharing: Improvements and Extensions. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1292–1303.
79. Bachiega, J.; Costa, B.; Carvalho, L.R.; Rosa, M.J.F.; Araujo, A. Computational Resource Allocation in Fog Computing: A Comprehensive Survey. *ACM Comput. Surv.* **2023**, *55*, 1–31. [CrossRef]

information

MDPI

Article

Privacy and Security Mechanisms for B2B Data Sharing: A Conceptual Framework

Wanying Li, Woon Kwan Tse * and Jiaqi Chen

Department of Information Systems, City University of Hong Kong, Hong Kong 999077, China; wanyinli8-c@my.cityu.edu.hk (W.L.); jchen2244-c@my.cityu.edu.hk (J.C.)
* Correspondence: iswktse@cityu.edu.hk

Abstract: In the age of digitalization, business-to-business (B2B) data sharing is becoming increasingly important, enabling organizations to collaborate and make informed decisions as well as simplifying operations and hopefully creating a cost-effective virtual value chain. This is crucial to the success of modern businesses, especially global business. However, this approach also comes with significant privacy and security challenges, thus requiring robust mechanisms to protect sensitive information. After analyzing the evolving status of B2B data sharing, the purpose of this study is to provide insights into the design of theoretical framework solutions for the field. This study adopts technologies including encryption, access control, data anonymization, and audit trails, with the common goal of striking a balance between facilitating data sharing and protecting data confidentiality as well as data integrity. In addition, emerging technologies such as homomorphic encryption, blockchain, and their applicability as well as advantages in the B2B data sharing environment are explored. The results of this study offer a new approach to managing complex data sharing between organizations, providing a strategic mix of traditional and innovative solutions to promote secure and efficient digital collaboration.

Keywords: data sharing; blockchain; B2B architecture; privacy; security

1. Introduction

In the era of big data, enterprises' data have become a strategic resource to promote global economic progress. The exchange and sharing of data for collaboration, informed decision-making, and sustainable operation are becoming increasingly important. Business-to-business (B2B) data sharing enables organizations to leverage the vast amount of information available, gaining valuable insights and driving innovation. Shared access to business data lays the foundation for service processes for companies within the same industry and even across industries and countries. According to a study by Gartner [1], sharing data externally by data and analytics leaders results in three times the measurable economic benefits compared to those who do not share data. In addition, from the value chain perspective, data sharing in this model has significant benefits, such as helping companies expand their business, improve operational efficiency, and improve customer service.

However, despite the profitable practice, there are still plenty of valid reasons for companies to keep their data closed and avoid sharing the data with third parties, with concerns around privacy, security, access control, and data governance. Traditional methods lack robust security, exposing sensitive business data to cyber threats and unauthorized access while posing significant privacy risks. Existing practices are inadequate in efficiently managing the diverse data storage architecture and increasing data volumes, leading to storage, retrieval, and overall data management difficulties. These inefficiencies hinder business operations and are compounded by a lack of transparency and traceability in auditing data transactions. Moreover, the rigidity and complexity of access control within

traditional systems fail to meet the dynamic needs of businesses, and there are substantial challenges in ensuring the integrity and consistency of data, especially in environments with frequent updates. Compliance with data sharing and privacy standards like General Data Protection Regulation (GDPR) adds another layer of complexity, requiring substantial resources and effort from businesses [2].

Recognizing the urgent need for privacy protection technologies and enhanced security protocols, this study aims to explore the privacy concepts and security mechanisms of B2B data sharing. Our main contribution is to propose a reference architecture for data sharing technologies that can be widely adopted based on enterprise infrastructure. It identifies fundamental security requirements and discusses potential solutions to achieve these goals while highlighting open challenges in this domain. The purpose of this paper is to present a conceptual framework through which we analyze and discuss our approach before conducting field tests. This study is expected to bridge the gap between current practices and emerging challenges in B2B data sharing, offering a comprehensive and viable solution that addresses current issues and lays a foundation for future research in this domain.

The rest of this paper is structured as follows: Section 2 presents the related work. Section 3 is about the significance of this research. Section 4 is the introduction to the theoretical basis of the paper. Sections 5–7 introduce the proposed data sharing system framework and analyze its characteristics. Finally, Section 8 summarizes and prospects the paper.

2. Related Work

The current landscape of B2B data sharing among organizations is complex and challenging. One of the primary challenges in B2B data sharing revolves around ensuring data security and privacy while maintaining accessibility and usability. Blockchain's integrity, immutability, decentralization, and verifiability are key strengths for data sharing in multi-party scenarios [3]. Some studies have explored the utilization of blockchain technology as a potential solution. J. Chi et al. [4] proposed a blockchain-based data sharing scheme, which combines identity authentication, Hyperledger Fabric, and community detection algorithms, effectively ensuring security and efficiency in industrial IoT data sharing. Xuan et al. [5] propose a dynamic incentive model for data sharing using blockchain with smart contracts, enabling trust and automated transactions between large numbers of users. A novel approach to data sharing was introduced by Al-Zahrani [6] in the form of a subscription-based model that utilizes the advanced features of blockchain technology. The proposed Data as a Service (DaaS) model has the potential to offer significant benefits to businesses seeking secure and efficient methods of data sharing. However, blockchain's limitations in scalability and its difficulty in storing vast amounts of data have been acknowledged as significant drawbacks [7].

To address these issues, alternative approaches such as off-chain storage solutions have been proposed. Cheng Xu et al. [8] propose SlimChain, a novel blockchain system that scales transactions through off-chain storage and parallel processing. Kete Wang et al. [9] propose a three-tier system architecture for efficient and transparent data sharing, storing hashing and response records on the blockchain and original data in an off-chain database, a concept that partially inspired our framework's architecture.

The migration of data to the cloud has become a common practice in recent times due to the advanced cloud infrastructure, which offers higher accessibility, lower response time, and more cost efficiency [10]. Despite the significant benefits of data sharing in cloud computing, the outsourcing of data deprives users of direct control, which gives rise to security concerns and poses challenges [11,12]. Kotha et al. [13] explored various problems and challenges of data sharing in cloud environments, highlighting the need for robust encryption and access control mechanisms. Song et al. [14] further examined and compared current encryption and key management techniques in cloud storage, providing a foundation for the encryption strategy employed in our framework.

Access control in B2B frameworks has been thoroughly investigated in works such as those by Gai et al. [15], who proposed a blockchain-based access control scheme using role-based access control (RBAC) specifically tailored for lightweight data sharing among different organizations. Xu et al. [16] proposed a blockchain-based secure data sharing platform with fine-grained access control (BSDS-FA) to prevent data leakage. Yang et al. [17] present a ciphertext-policy attribute-based conditional proxy re-encryption (CPRE) scheme, which allows efficient user revocation and resource-constrained devices to access cloud data. The granular access control in our framework takes cues from these studies, aiming to provide fine-grained permissions management.

Various advanced technologies have been used to facilitate secure data sharing. Proxy re-encryption as a tool for secure data sharing has been gaining traction. The survey by Qin et al. [18] laid the groundwork for using proxy re-encryption in secure data sharing, which has been adapted in our framework to streamline access delegation. Additionally, the application of homomorphic encryption in data sharing has been explored by researchers like Zhu et al. [19], whose insights into practical homomorphic encryption applications have influenced the privacy-preserving aspects of our framework.

Despite the extensive research on data exchange mechanisms, there is still a significant gap in the literature regarding a comprehensive theoretical framework that covers data storage, security, privacy protection, etc. Many existing studies only focus on specific aspects of B2B data sharing and do not consider the need for an integrated approach that addresses the interconnected nature of these components. Therefore, there is an urgent need for a theoretical framework that provides a comprehensive solution to the multifaceted challenges inherent in B2B data sharing, facilitating the development of more robust and effective solutions.

3. Research Significance

In this study, after analyzing the need to have a secure mechanism in B2B data sharing, our most significant contribution is to propose a bullet-proof three-layer B2B data sharing framework while maintaining a reasonable degree of accessibility. It is based on blockchain technology, offering a comprehensive solution to the diverse challenges in B2B data sharing. This framework comprises a data storage layer, a privacy preservation layer, and an access control layer. This layered approach is designed to provide a comprehensive solution that balances the need for data accessibility with security and privacy requirements. The framework is grounded in blockchain technology, integrating emerging technologies such as distributed hash table, homomorphic encryption, and proxy re-encryption. The proposed conceptual framework has the following features:

- Scalability and adaptability—a scalable and adaptable framework is essential with increasing data volumes. Blockchain combined with off-chain storage solutions caters to this need, enabling efficient management of large datasets without compromising the blockchain's performance [20].
- Flexibility in storage architectures—cloud storage offers advantages such as lower cost, metered service, scalable, and ubiquitous access, but raises concerns about data integrity and privacy [21]. Our framework offers a variety of off-chain storage options, allowing businesses to decide on their storage architectures flexibly based on their needs and security considerations.
- Efficient data location and retrieval—the integration of distributed hash tables (DHTs) ensures efficient and secure access to data across distributed environments, addressing the challenges of data storage and retrieval in a diverse storage landscape [22].
- Privacy preservation—utilizing encryption and privacy-enhancing technologies like data masking ensures that sensitive data remain confidential, catering to the increasing privacy concerns in data sharing [23].
- Efficient access rights control—the framework's capability to manage access rights efficiently without requiring extensive re-encryption processes enhances its practicality and efficiency, particularly in dynamic business environments.

- Compliance and auditability—the blockchain component of the framework provides a transparent and auditable record of all data sharing requests and access events, enhancing accountability and regulatory compliance [24].

4. Preliminaries

4.1. Data Encryption Technology

4.1.1. Additive Homomorphic Encryption

Homomorphic encryption is a cryptographic technique that enables computations on encrypted data without decryption. It allows computation on encrypted data, yielding the same results as unencrypted data, and is faster and consumes less memory space compared to traditional methods [25]. In other words, it will enable operations to be performed on data while still in encrypted form. In 1999, French cryptographer Paillier [26] published the Paillier algorithm at Eurocrypt, one of the most prestigious academic conferences in cryptography. This became the initial source of additive homomorphic encryption algorithms. The algorithm implemented in this system is rooted in the cryptographic technology that harnesses computational complexity theory to solve mathematical problems. It is the only additive homomorphic encryption algorithm designated by the ISO homomorphic encryption international standard [27]. It offers a range of desirable properties, including the ability to perform secure computations on encrypted data while maintaining the privacy and integrity of the data. It can satisfy the following properties:

$$\mathrm{E(x)} \oplus \mathrm{E(y)} = \mathrm{E(x+y)} \tag{1}$$

- Randomly select two large prime numbers of equal length p, q.

n and λ are defined by $n = pq$, $\lambda = lcm(p-1,\ q-1)$, where $lcm(x,\ y)$ denotes the least common multiple of x and y.

- Randomly select integer $g \in Z_{n^2}^*$, $L(x) = \frac{x-1}{n}$ is defined, calculate $\mu = \left(L\left(g^{\lambda} mod\ n^2\right)\right)^{-1} mod\ n$, and generate keys as

$$\begin{cases} public-key = (n, g) \\ private-key = (\lambda, \mu) \end{cases} \tag{2}$$

- To perform encryption, input plaintext information m, $0 \le m \le n$, select random numbers $r \in Z_n^*, gcd(r, n) = 1$, and compute the encrypted ciphertext:

$$c = g^m r^n mod\ n^2 \tag{3}$$

- To perform decryption, compute the plaintext message:

$$m = L(c^{\lambda}\ mod\ n^2) \cdot \mu\ mod\ n \tag{4}$$

This is particularly useful for protecting data privacy in scenarios where data need to be handled in a secure and confidential manner (e.g., in cloud computing or secure data analytics) without exposing the plaintext.

Homomorphic encryption provides a significant advantage over traditional encryption methods. Unlike traditional encryption, which requires decryption before further operations, homomorphic encryption allows computations to be directly executed on encrypted data, eliminating the need for repeated encryption decryption [28]. It not only strengthens data security by reducing exposure to potential breaches during computation but also facilitates secure and privacy-preserving data processing in distributed environments, such as cloud computing and outsourced data analysis. Additionally, homomorphic encryption enables secure collaboration on sensitive data across different parties without compromising confidentiality, making it a powerful tool for advancing privacy-preserving technologies in various domains, including healthcare, finance, and enterprise data sharing [29].

4.1.2. Proxy Re-Encryption

Proxy re-encryption (PRE) is a cryptographic technique allowing an entity (called a proxy) to re-encrypt encrypted data from one key to another without accessing plaintext information. Proxy re-encryption was first introduced by Matt Blaze et al. [30] at the 1998 Eurocrypt conference. Their work laid the groundwork for the development of encryption schemes that allow a proxy to transform ciphertexts from one key to another, without learning anything about the underlying plaintext.

This technique has applications in various ways, especially in protecting data privacy and enabling secure data sharing and cloud storage. Such a process allows the data owner to delegate to an agent and have the agent pass the data between them while keeping the data encrypted. The critical steps in proxy re-encryption are as follows:

- Key generation: Each entity has its own public key(pk) and private key(sk). The proxy that performs the re-encryption does not need to decrypt the data but can directly transform ciphertexts from being decryptable. For example, the process whereby entity A generates a re-encryption key using the public key of itself and the private key of B can be represented as follows:

$$rk_{\mathrm{A}\to\mathrm{B}} = f(sk_{\mathrm{A}}, pk_{\mathrm{B}}) \tag{5}$$

- Encryption: the data owner encrypts the data m using the agent's public key and passes the ciphertext to the agent.

$$c_A = Enc_{pkA}(m) \tag{6}$$

- Proxy re-encryption: The proxy re-encrypts the data to another key using its private key, without knowing the plaintext. In this way, the proxy obtains a new ciphertext and keeps the data encrypted.

$$c_B = ReEnc_{rk_{A\to B}}(c_A) \tag{7}$$

- Decryption: The recipient of the data decrypts the agent's re-encrypted ciphertext using its private key to obtain the final plaintext. In the previous example, B can use its private key to decrypt

$$c_B : m = Dec_{sk_B}(c_B) \tag{8}$$

The design and application of proxy re-encryption relies on specific scenarios and requirements to ensure privacy protection and security in data sharing and delegation contexts.

4.1.3. Data Perturbation

Data perturbation is a privacy-preserving technique that aims to protect the privacy of sensitive information while maintaining the validity and usefulness of the data by transforming or perturbing the original data with a certain degree of modification.

The concept of data perturbation dates back to methods developed in statistical disclosure control (SDC), which were introduced to protect privacy in statistical databases. Early forms of data perturbation were simple and aimed at adding "noise" to data to prevent the identification of individuals from published datasets. Dalenius introduced various methods to assess and mitigate the risk of disclosing information in statistical databases [31]. After that, more sophisticated methods were developed. The methods include data swapping and adding random noise aimed at protecting individual data while allowing for statistical analysis [32].

This technique is now usually applied in scenarios where sensitive data need to be shared or processed to prevent unauthorized access or disclosure. Commonly used data masking techniques include methods such as desensitization, where a portion of the information in sensitive data is replaced with fuzzy or fictitious values.

Dwork [33] introduced the concept of differential privacy, which revolutionized the field of data privacy. It often uses perturbation techniques, such as adding Laplacian and Gaussian noise to the data, offering strong guarantees against various types of inference attacks. Some other methods include noise addition, generalization, erasure, synthetic data generation, etc.

4.2. Data Storage Technology

4.2.1. Blockchain

Blockchain is a distributed and immutable digital ledger technology. The initial research on this technology came from Nakamoto's [34] study on Bitcoin transactions, A Peer-to-Peer Electronic Cash System, which is also the first paper to introduce Bitcoin. It is often associated with cryptocurrencies such as Bitcoin, but its applications go far beyond that. Essentially, blockchain is a decentralized database that records transactions on a computer network securely and tamper-proof. When data users cannot fully trust centralized institutions to store and process data, blockchain has more potential to guarantee data access, control, and security. Over time, it has been realized that this technology is applicable to digital currencies and can have far-reaching implications in many other areas. The technology can be applied in various areas, including supply chain management, voting systems, and protection of digital assets to ensure consistency in transmission protocols, data ownership, and other elements.

Blockchain is usually classified as a public blockchain, private blockchain, or permitted blockchain, depending on the participants:

- Public blockchain: Bitcoin and Ethereum are two famous examples of a public blockchain whose data are completely open and can be accessed and queried by anyone.
- Private blockchain: private blockchains are typically used within an organization as they restrict access to the blockchain, usually to authorized users or entities.
- Permissioned blockchain: In comparison, the permissioned blockchain is somewhere in between, managed by a group of organizations or entities for more flexible, secure, and customizable applications. It is suitable for scenarios where multiple companies collaborate. This hybrid approach enables interaction with other entities while meeting privacy and authorization needs. Permissioned blockchains can use more streamlined consensus mechanisms such as Practical Byzantine Fault Tolerance (PBFT), Raft, or Hybrid Consensus.

4.2.2. Distributed Hash Table

A distributed hash table (DHT), also known as the Kademlia algorithm, is a type of distributed computing system. The concept of distributed hash tables was introduced by four seminal papers that were published in quick succession in the early 2000s, which proposed the architectures of Chord, CAN, Pastry, and Tapestry [35–38]. Each of these works contributed to the foundational understanding and development of DHTs used in decentralized applications today. For example, a recent study shows that it can be combined with encryption technology to provide secure data storage and retrieval and be used to manage decentralized identities and access control [39].

A DHT aims to enable data to be stored and retrieved efficiently. In terms of data sharing, using a DHT to build distributed data storage systems can ensure that data are evenly distributed among participants, improving data availability and fault tolerance. A DHT is based on the concept of hash function, which distributes data to multiple nodes and locates and retrieves the stored data through hash values. Each node is responsible for maintaining a portion of the data and works together through network protocols to present the entire system with a consistent, distributed hash table structure. The basic theory of a DHT includes the following key concepts:

Hash functions and hash rings: a hash function is a mathematical function that maps an input to a fixed-length output. For a DHT, a hash function is typically used to map the

identity of the data to a unique hash value, which is then mapped to a ring space, called a hash ring. The hash value determines which node in the DHT the data should be stored on, and each node occupies a position in the hash ring. The mathematical representation of the hash function is

$$H(x) : x \rightarrow [0, 2^n - 1] \tag{9}$$

where n is the number of bits in the hash output and the SHA-1 hash function is used in the Kademlia DHT.

Data storage and retrieval: To store or retrieve data in the DHT, the hash value of the data identifier is first calculated by the hash function. Then, the nearest node to this hash value is found on the hash ring. The data are stored on this node or retrieved from this node. This approach makes the data evenly distributed among the nodes. The mathematical representation can be to find the node that satisfies the following conditions, where k is the identifier of the node:

$$Node = \min(|H(k) - H(id)|) \tag{10}$$

Node communication protocol: some kind of communication protocol is required between the nodes so that they can work together to store and retrieve data. A basic message format is $Message = \{\text{Type}, \text{SourceID}, \text{TargetID}, \text{Payload}\}$.

Overall, a DHT also considers features such as distributed consistency and fault tolerance, which helps to achieve the efficient storage and retrieval of data in a distributed environment.

5. Framework Overview

In this work, we propose a three-layer blockchain-based framework for B2B data sharing. As illustrated in Figure 1, the structure comprises a data storage layer for secure data handling, a privacy preservation layer ensuring data privacy through advanced encryption, and an access control layer for dynamic permission management, offering a comprehensive solution for secure and efficient business data exchange.

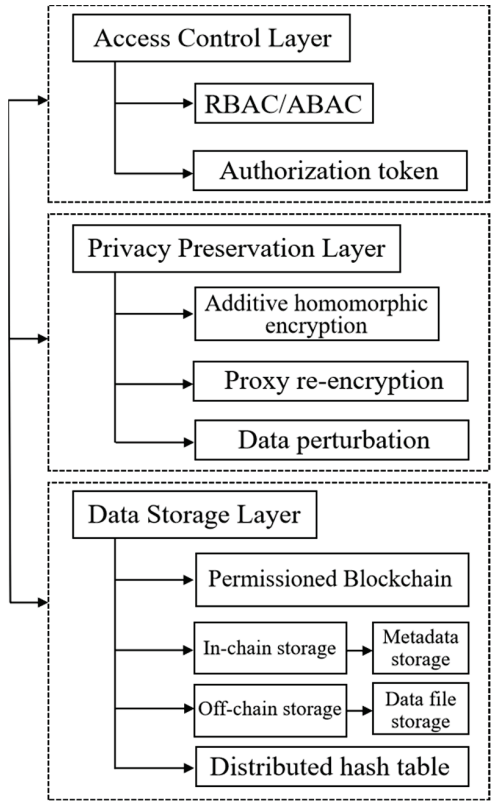

Figure 1. Three-layer B2B data sharing framework. RBAC and ABAC in the figure are abbreviations for role-based access control and attribute-based access control.

- Access Control Layer

In the access control layer, the data owner has the flexibility to control and adjust user access to data, including options to revoke or reassign access rights. This layer employs different access control mechanisms for local and cloud storage. For local and cloud storage, we employ a traditional method such as role-based access control (RBAC). For cloud storage, we employ a complex approach for access control, utilizing an authorization token for each data user.

- Privacy Preservation Layer

The privacy preservation layer incorporates integrating additive homomorphic encryption, data perturbation, and proxy re-encryption mechanisms. The workflow of the privacy preservation layer primarily involves two critical stages: key generation and data encryption.

In the key generation stage, the data owner creates unique public and private key pairs for themselves as well as for each authorized user using asymmetric encryption algorithms, and securely transmits the secret keys to the corresponding users. Furthermore, a proxy re-encryption key will be created for each authorized user, which will be used later.

In the data encryption stage, we employ different encryption approaches for local and cloud storage. For local storage, the privacy layer emphasizes secure data storage on the device and secure transmission within the local network, employing traditional encryption methods that combine symmetric encryption with asymmetric encryption key exchange. In the context of cloud storage, the focus shifts more towards the security and management of data within the cloud environment, considering the unique characteristics and challenges posed by cloud services. Our solution ensures that data remain secure and properly managed, regardless of whether they are stored locally or in the cloud.

- Data Storage Layer

In our B2B data sharing platform design, we adopted a hybrid approach to data storage, combining a permissioned blockchain with off-chain databases. To ensure data integrity and transparency, the permissioned blockchain stores metadata, while business data are offloaded to off-chain storage systems. We use distributed hash tables (DHTs) for content addressing, facilitating the efficient location and retrieval of files across diverse storage formats.

6. Data Sharing Process

In this section, we will present the complete data sharing process under the framework proposed in Section 5. The complete process is as follows (shown as Figure 2):

Figure 2. Data sharing process.

Step 1: Data Storage:

Data are strategically stored based on their nature and sensitivity. Essential information such as metadata are stored on-chain for transparency and auditability. Bulk business data, which are typically voluminous, are stored off-chain in various storage systems like databases and cloud storage solutions for privacy and efficiency. More details will be provided in Section 7.

Step 2: Data Encryption:

Before the data are uploaded, they are encrypted using additive homomorphic encryption. This allows certain calculations to be performed on the encrypted data without needing to decrypt them. To further enhance privacy, data perturbation is applied to the

datasets intended for sharing. This modifies the data slightly to mask the actual properties and values of the original data while maintaining overall statistical characteristics.

Step 3: Access Control:

The data owner can have fine-grained control over the sharing of data, which includes granting access, revoking it, and reassigning it. In the proposed framework, the data owner can revoke access for specific users and assign different access rights, without the need for key redistribution or data re-encryption.

Step 4: Data Access:

When a data request is made, the distributed hash table (DHT) is utilized to locate the requested file across the diverse storage landscape. The system then verifies the requester's authorization to access the data. If authorized, proxy re-encryption is employed to transform the data, which are encrypted for the owner's key, into a format that the requester's key can decrypt, facilitating secure data sharing without exposing the actual data.

Step 5: Transaction Audit:

Every action on the platform, including data sharing requests and access events, is logged on the blockchain for auditability.

7. Analysis and Discussion

In this chapter, we will provide a comprehensive explanation of the three-layer B2B data sharing framework that was proposed in Section 6. We will critically analyze each of its components and mechanisms to demonstrate its effectiveness. The following sections will break down these layers, present a detailed analysis of their functionalities and interdependencies, and illustrate how, collectively, they form a cohesive and potent solution to the complex challenges of secure and efficient data sharing in business-to-business contexts.

7.1. Data Storage Layer

In the data storage layer, we use a hybrid approach of a permissioned blockchain with off-chain databases, where metadata are stored on blockchain for transparency, and business data are stored off-chain for efficiency. A DHT is used for efficient file location and retrieval. In this section, we will discuss how data are strategically stored based on their nature and sensitivity and explain the significance of a hybrid storage approach in terms of transparency, auditability, privacy, and efficiency.

- Permissioned Blockchain

Blockchain technology offers two distinct types: open, permission-less systems like Bitcoin or Ethereum; and controlled, permissioned networks such as the Hyperledger Project from The Linux Foundation. In a permission-less or public blockchain, the identities of the participants in the system remain anonymous. It implies that anyone can join or leave the blockchain network at any time, increasing the risk of security breaches in the network. This is quite dangerous for B2B data sharing. Nevertheless, in a permissioned or private blockchain, only a known and identifiable set of participants are explicitly admitted to the blockchain network. It reduces the presence of malicious actors within the network. Consequently, only authenticated and authorized actors can participate in the network, which enhances the security of the system, as required by enterprise applications. This is important to the accessibility of B2B data sharing. Based on the above arguments, we choose to base our framework on a permissioned blockchain for controlled access, enhanced security, and regulatory compliance.

- In-chain Storage

The main purpose of in-chain storage is to record all critical data operations, including data creation, modification, and access events, as well as the hash value of the data themselves. It takes advantage of the fundamental feature of blockchain, which is that data cannot be changed once they are recorded, ensuring the credibility and consistency of the information.

When new data are created and uploaded to the platform, their metadata (e.g., file size, format, and time of creation) and the hash value of the data are recorded on the blockchain.

The hash value is generated by applying a cryptographic hash function to the data content, which provides a unique digital fingerprint of the data. This hash value plays a key role in the data lifecycle and is used to verify that the data have not been tampered with. If the data are modified, the specific details of the modification (including the modification time and the identity of the modifier) and the hash value after modification will be recorded. In this way, any change to the data will leave an immutable trace, guaranteeing the historical traceability of the data. This is important to the data integrity of B2B data sharing. Whenever data are accessed, the time of access, the identity of the visitor, and the nature of the access (e.g., viewed or downloaded) are recorded. This provides transparency of data usage and facilitates monitoring and auditing. A standard metadata will have the following structure (Figure 3):

```
Record{
  TransactionID: Unique identifier of the transaction,
  OperationType: Type of operation, 'Create', 'Modify', or 'Access'
  Timestamp: Time when the operation was performed,
  Metadata:{
    Identifier: Unique identifier of the dataset,
    Size: Size of the dataset,
    Format: Format of the dataset (e.g., CSV, JSON),
    Owner: Identity of the entity uploading the dataset
  }
  DataHash: Cryptographic hash of the dataset,
  Operator: Identity of the entity uploading the dataset,
  OperationDetails: Specific details of the operation (e.g., 'view', 'download')
}
```

Figure 3. Standard metadata structure.

- Off-chain Storage

In our B2B data sharing framework, off-chain storage is a key part of managing large and complex business data, especially when it comes to traditional databases and cloud storage solutions. These storage options are suitable for real-world business data, such as large datasets and bulky files, which are often not suitable for storage on the blockchain due to scalability limitations. Thus, off-chain storage provides the necessary capacity and flexibility to efficiently handle large datasets while addressing the limitations of the blockchain in handling high volume data storage. The combination of using in-chain and off-chain storage is to balance the need of robustness deployed by in-chain storage while balancing the scalability concern of voluminous B2B data deployed by off-chain storage.

Cloud storage is a popular choice for many businesses facing growing data volumes and large storage requirements. It reduces the cost of database construction and maintenance and facilitates data exchange and circulation. However, for companies highly sensitive about data privacy, storing substantial data with centralized cloud service providers (CSPs) can pose significant risks. Storing large amounts of data with semi-trusted third parties could lead to misuse and disclosure of proprietary information.

Therefore, for real-time or sensitive data, businesses often prefer local storage on their own devices, exposing access interfaces to the sharing systems. This approach provides companies with more direct control over their data, reducing the risks associated with third-party data breaches. Local storage ensures the immediacy and sensitivity of the data while allowing for safe sharing and access when needed. In addition, off-chain storage systems offer flexibility to comply with various regulatory requirements, such as GDPR, especially when sharing data that could potentially identify individuals. They can be

customized to align with diverse compliance standards, an essential aspect for businesses operating across different legal landscapes.

Our framework offers a variety of off-chain storage choices, allowing businesses to flexibly decide on their data storage method based on their needs and security considerations. Whether choosing cloud storage for its cost-effectiveness and convenience or opting for local storage for greater security and control, our framework supports effective and secure data sharing while ensuring data safety and privacy.

- Distributed hash tables

In the data storage layer, the DHT serves to locate datasets across different business entities. To illustrate how the DHT functions in the framework, consider an example of supply chain management. In such a scenario, the implementation of a DHT facilitates a decentralized approach to sharing data among various entities such as manufacturers, suppliers, distributors, and retailers. Each of these businesses operates as a node within the DHT network, representing their own unique databases. These nodes could be servers or data centers belonging to different organizations participating in the data sharing process. In addition to storing actual data, these nodes keep the DHT to hold and share information about where and how specific data can be accessed.

For each business's database, a unique key is generated using a hash function. This key is typically derived from identifiable attributes of the business, i.e., identifier for the dataset mentioned above. When a business decides to share specific data, like inventory levels or production schedules, it does not directly store these data in the DHT. Instead, it registers an entry in the DHT under the generated key, with the entry containing metadata or pointers indicating how and where these data can be accessed, such as an API endpoint or a database query interface. A DHT entry may look like Figure 4:

```
Key:ef6523a34d8e4a018b14a7abf1c79f59a8c4ef15`
Value:
- Database Access Point: `http://db.tech6523.com`
- API Endpoint: `https://api.tech6523.com/inventory`
- Metadata: `Inventory data, last updated on 2023-11-18`
```

Figure 4. A DHT entry sample.

When a user wants to find a database, they use the hash function on the dataset's identifier to generate the key and query the DHT. The DHT responds by providing the information about where the database is located.

It should be noticed that every node in the distributed network has a part of the DHT, rather than the entire table duplicated in all nodes. When there is a request for the corresponding value for the key, a peer receives the request and checks for the key in its own table. If it is available, the value will be returned or else the request will be passed on to the peers until the value is found. This is important in B2B data sharing because the use of a DHT can prevent the problem of putting all eggs in one basket.

7.2. Privacy Preservation Layer

The privacy preservation layer incorporates integrating additive homomorphic encryption, data perturbation, and proxy re-encryption mechanisms. The workflow of the privacy preservation layer primarily involves two critical stages: key generation and data encryption.

7.2.1. Key Generation

Based on asymmetric encryption algorithms, the data owner generates unique public and private key pairs for each authorized user. The first one is the master key pair for

the data owner (pk_a, sk_a), where pk_a is kept secret. The second key public/private pair (pk_b, sk_b) is created for each authorized user. Then, the data owner securely sends the corresponding key pair to the authorized user, which can be realized by a temporary key exchange protocol or existing public/private key pair of both. Both the data owner's public key pk_a and the authorized user's public key pk_b are considered public. Furthermore, for each authorized user, the data owner creates a proxy re-encryption key $rk_{a\to b}$, which is used to secure data sharing between different users without decryption and re-encryption. The key is generated using sk_a and pk_b, that is

$$rk_{a\to b} = Generate\ Re\ EncrytionKey(sk_a, pk_b) \tag{11}$$

7.2.2. Data Encryption

There are different encryption approaches for local storage and cloud storage based on their different storage environments.

- Local storage:

For local storage, the privacy layer focuses on the secure storage of data on the device and the secure transmission within the local network. Compared to cloud storage, local storage is less exposed to external threats such as data breaches happening in cloud service providers. Therefore, we employ traditional encryption methods combining symmetric encryption with the asymmetric encryption key exchange method. This mechanism is well established and therefore will not be further discussed here.

- Cloud storage:

For cloud storage, the privacy layer focuses more on the security and management of data in the cloud environment, while considering the characteristics and challenges of cloud services.

If you consider the cloud service provider as a semi-trusted third party, there are potential security issues. One of the main security concerns in cloud computing is how the data are being used by a third-party cloud service provider. Combining access control with re-encryption is a viable solution for protecting data from unauthorized access and cloud breaches [40]. In this framework, we adopt homomorphic encryption and data perturbation to protect sensitive data stored on the cloud from unauthorized access.

Before the data are uploaded to the cloud platform for storage, the data owner first implements data masking on each data record, replacing, disrupting, or partially hiding sensitive information to ensure that the original content of the data has been protected when they leave the local environment. Then, they are further encrypted using homomorphic encryption. For each authorized user, corresponding authorization tokens are generated. These tokens determine the user's level of access to the encrypted data. More details about the authorization process will be provided at the access control layer. Algorithm 1 illustrates the encryption process, which involves data masking and encryption:

Algorithm 1 Data Masking and Encryption

procedure MASKANDENCRYPT (d, pk_a)
  $n \leftarrow$ number of attributes in d
  $m \leftarrow$selectRandomNonNegativeNumber()
  $r \leftarrow$ generateRandomIntegerList$(n+m)$
for $i \leftarrow 1$ to $n+m$ **do**
  $d'_i \leftarrow d_i + r_i$
  $E_{pk_a}(d'_i) \leftarrow EncryptAlgorithm(d'_i, pk_a)$
end for
  return $\left\{E_{pk_a}(d'_1), E_{pk_a}(d'_2), ..., E_{pk_a}(d'_{n+m})\right\}$
end procedure

Assume that the data owner's data consists of multiple records, and each data record d contains n attributes. Each record can be represented as $\{d_1, d_2, \ldots, d_n\}$, where d_i is the value of the i^{th} attribute. In the process, we need to mask the attribute first. To protect sensitive data, the owner can mask the attributes and their values by selecting n random integers, denoted $\{r_1, r_2, \ldots, r_{n+m}\}$. These integers are chosen from a specific number field. The value of m is an additional security parameter to mask the actual number of attributes of the data record, which is also randomly selected and may vary from record to record based on the security requirements of the data owner for the given application. For each data record d_i, data owner combines it with the correspond random number r_i, and then encrypted it using the public key pk_a, generating $E_{pk_a}(d_i + r_i)$.

The core of this process is to increase the privacy of the data through random numbers, so that the privacy of the data content is still protected even if the data are stored on an external cloud service. Through this mechanism, even if someone accesses these data, they will not be able to determine the original values of the attributes. Additionally, they will not be able to determine the total number of attributes that existed in the original data record. This enhances the protection of not only the raw attribute values but also the data structure information. This is important to the privacy aspect of B2B data sharing.

7.3. Access Control Layer

In the access control layer, the data owner has the flexibility to control and adjust user access to data. This includes the option to revoke a user's access rights or to reassign them in the future. Additionally, the use of data masks ensures that even if a user's access is revoked, the data they previously accessed cannot reveal any confidential information.

It is similar to the privacy preservation layer, as different access control mechanisms are employed for local storage and cloud storage. It is under direct physical control and therefore we can use traditional access control mechanisms such as role-based access control (RBAC). For finer-grained control, attribute-based access control (ABAC) can be used, where user attributes and environmental factors can be considered.

However, the situation becomes complicated for cloud storage because of privacy, security, and regulation issues. Data must be encrypted before they are uploaded to the cloud service provider, which is generally a secure practice, but it may introduce certain challenges such as a performance overhead and access control complexity. This is quite normal in B2B data sharing. Before the authorized user obtains the data, the data owner has to download and decrypt the encrypted data from the CSP first and re-encrypt the data using the data user's private key and finally upload to the CSP again. It can introduce a performance overhead, especially in decryption and encryption for large datasets. Moreover, managing who can access which pieces of data can become complex, especially in a dynamic B2B environment where access needs may frequently change. This requires a robust access control mechanism that integrates well with the encryption system.

To address these challenges, we adopt a proxy re-encryption mechanism. We generate authorization tokens for the encrypted data $E_{pk_a}(d + r)$ to manage the access control. Consider four stages: access grant, data access, access revocation, and access reassign.

7.3.1. Access Grant

If data owner wants to grant access to the data user for a set of attributes S (where is S a subset of d), an authorization token T_b^d is generated for the user corresponding to the attributes, that is

$$T_b^d = \left\{ DataUser,\ \ rk_{a\to b}, \left\{ E_{pk_b}(\alpha_1), ..., E_{pk_b}(\alpha_{n+m}) \right\} \right\} \tag{12}$$

where

$$\alpha_i = \begin{cases} -r_i & if\ 1 \le i \le n \wedge d_i \in S \\ -d_i' & otherwise \end{cases} \tag{13}$$

$rk_{a\to b}$ is the proxy re-encryption key created in the privacy preservation layer. d_i' is calculated in Algorithm 1. For $1 \le i \le n$, $d_i' = d_i + r_i$; for $n < i \le n+m$, $d_i' = r_i$.

If the data owner does not want the data user to access d, which means $S = \varnothing$, she does not generate any authorization token, effectively setting the data user's token $T_b^d \leftarrow null$.

The data owner creates a list of authorization tokens that specifies which users can access which parts of the data. This list, denoted as T^d, includes tokens for accessing either the entire record or specific parts of it, depending on the permissions granted. If a user's token for a particular data record is null, indicating no access, it is not included in the list. Finally, the data owner uploads the authorization tokens T^d and the corresponding encrypted data $E_{pk_a}(d')$ records to the cloud.

7.3.2. Data Access

When the data user wants to access the data, he/she will send a data request to the cloud. When receiving a data request from the data user, the cloud first checks if there is an entry for the data user in an authorization token list T^d. If no entry is found, the process is aborted. If the data user is found to be authorized, the cloud proceeds with a multi-step re-encryption and decryption procedure. This process utilizes the proxy re-encryption key $rk_{a\to b}$ to convert $E_{pk_a}(d')$. The cloud then performs an additive homomorphic calculation on the encrypted data as follows:

$$E_{pk_b}\left(d_i{}' + \alpha_i\right) \leftarrow E_{pk_b}\left(d_i{}'\right) +_h E_{pk_b}(\alpha_i),\ \ 1 \le i \le n+m \tag{14}$$

Noted that $E_{pk_b}(\alpha_i)$ is from the user's authorization token T_b^d, and $+_h$ is the additive homomorphic calculation. The result will be sent to data user, and then he/she can perform $D_{pk_b}\left(E_{pk_b}\left(d_i{}' + \alpha_i\right)\right)$ to get $d_i{}' + \alpha_i$ with private key sk_b, which is received securely from data owner in Privacy Preservation Layer. Algorithm 2 shows the data access process using homomorphic encryption, which is performed by the cloud.

Algorithm 2 Data Access with Homomorphic Encryption

procedure DATAACCESS $\left(E_{pk_a}(d'), rk_{a\to b}, T_d^b\right)$
  Check if an authorization token T_d^b exists for the data user
  if No entry for data user in T_d **then**
    return Access Denied
  else
    for $i \leftarrow 1$ to $n+m$ **do**
      $E_{pk_b}\left(d_i'\right) \leftarrow PRE.ReEnc\left(E_{pk_a}, rk_{a\leftarrow b}\right)$ where α_i is from T_d^b
      $E_{pk_b}\left(d_i' + \alpha_i\right) \leftarrow E_{pk_b}\left(d_i'\right) +_h E_{pk_b}(\alpha_i)$
    end for
    Send the result to data user
  end if
end procedure

The data user's decryption will yield the correct data only for the attributes he is authorized to access. The calculation of $d_i{}' + \alpha_i$ obtains d_i only if the data user has access to the i^{th} attribute of d. Unauthorized attributes will result in a value of 0 upon decryption. The additive homomorphic property ensures that the user can compute sums of the encrypted values without access to the private keys that encrypted them.

7.3.3. Access Revocation

The data owner can revoke a user's access to a data record d by removing the authorization token T_b^d corresponding to $E_{pk_b}(d')$. Considering that the cloud as a semi-trusted third party may not follow the rules of the protocol, we use a smart contracts mechanism

to enforce the execution of the token revocation. Such access revocation is to ensure the proper rights for accessing the assigned data in B2B data sharing.

7.3.4. Access Reassign

When the data owner decides to reassign a user's access to a data record d^1, after a previous access revocation, he/she generates a new authorization token that specifies the set of data attributes the data user is allowed to access. This set can either be the same as the user had access to before the access rights were revoked or a different set of attributes within the data record. The data owner then sends this token to the cloud, directing it to update the authorization list T^d associated with the data record d^1. It is similar to access revocation, with the aim of restoring proper data access rights in B2B data sharing.

8. Conclusions and Future Work

Data sharing between businesses can create value for all relevant parties. Each business needs to be clear when considering the framework for the selected data sharing, which is that although sharing will bring challenges to all parties, such as data privacy and access security issues, the expected benefits from sharing need to exceed the construction costs and risks of the sharing mechanism. This study provides a comprehensive solution to the privacy and security challenges of data exchange between businesses in the digital age by establishing a reliable three-layer B2B data sharing framework based on blockchain. Firstly, an in-depth analysis of the development trend of B2B data sharing is conducted, during which the basic structure and additional security and privacy requirements of data sharing are identified, emphasizing the urgency of privacy protection technologies and enhanced security protocols. On this basis, a framework including a data storage layer, privacy protection layer, and access control layer is designed to ensure secure and efficient business data exchange. The framework uses advanced encryption technology, ensures data privacy through compliance, and achieves fine control of data through dynamic permission management. The application of this innovative technology is expected to provide additional value to businesses in addition to ensuring the data sharing process.

Through in-depth research, it is believed that this framework can balance the confidentiality of data while achieving data sharing, providing a viable solution for businesses. Emerging technologies, such as homomorphic encryption and blockchain, their applications, and advantages in the field of B2B data sharing, are also explored. This paper deepens the understanding of privacy and security mechanisms in B2B data sharing, providing valuable insights for businesses participating in the digitalization process.

Overall, the blockchain-based B2B data sharing framework is expected to become an important support tool in the digital transformation of enterprises in the future, promoting more secure and efficient information circulation between industries. To improve and expand this data sharing framework, further research and practical implementation will be encouraged using various emerging technologies. In future, we plan to test the framework with real-time data to evaluate its performance based on the processing time and computation cost. We will conduct a comparative study using different encryption algorithms under different storage architectures and analyze the results as an extension of the existing work.

Author Contributions: Conceptualization, W.L., W.K.T. and J.C.; Methodology, W.L. and J.C.; Software, W.L. and J.C.; Validation, W.L. and J.C.; Formal analysis, W.L. and J.C.; Investigation, W.L. and J.C.; Resources, W.L. and J.C.; Data curation, W.L. and J.C.; Writing—original draft, W.L. and J.C.; Writing—review & editing, W.L., W.K.T. and J.C.; Visualization, W.L.; Supervision, W.K.T.; Project administration, W.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goasduff, L. *Data Sharing is a Business Necessity to Accelerate Digital Business*; The Gartner Group: Stamford, CT, USA, 2020; Volume 11.
2. Lee, W.S.; John, A.; Hsu, H.C.; Hsiung, P.A. SPChain: A Smart and Private Blockchain-Enabled Framework for Combining GDPR-Compliant Digital Assets Management With AI Models. *IEEE Access* **2022**, *10*, 130424–130443. [CrossRef]
3. Shen, M.; Zhu, L.; Xu, K. Blockchain and Data Sharing. In *Blockchain: Empowering Secure Data Sharing*; Shen, M., Zhu, L., Xu, K., Eds.; Springer: Singapore, 2020; pp. 15–27.
4. Chi, J.; Li, Y.; Huang, J.; Liu, J.; Jin, Y.; Chen, C.; Qiu, T. A secure and efficient data sharing scheme based on blockchain in industrial Internet of Things. *J. Netw. Comput. Appl.* **2020**, *167*, 102710. [CrossRef]
5. Xuan, S.; Zheng, L.; Chung, I.; Wang, W.; Man, D.; Du, X.; Yang, W.; Guizani, M. An incentive mechanism for data sharing based on blockchain with smart contracts. *Comput. Electr. Eng.* **2020**, *83*, 106587. [CrossRef]
6. Al-Zahrani, F.A. Subscription-Based Data-Sharing Model Using Blockchain and Data as a Service. *IEEE Access* **2020**, *8*, 115966–115981. [CrossRef]
7. Wei, Q.; Shen, Z. Improving Blockchain Scalability from Storage Perspective. In Proceedings of the ACM Turing Award Celebration Conference—China 2023, Wuhan, China, 28–30 July 2023; pp. 112–113.
8. Xu, C.; Zhang, C.; Xu, J.; Pei, J. SlimChain: Scaling blockchain transactions through off-chain storage and parallel processing. *Proc. VLDB Endow.* **2021**, *14*, 2314–2326. [CrossRef]
9. Wang, K.; Yan, Y.; Guo, S.; Wei, X.; Shao, S. On-Chain and Off-Chain Collaborative Management System Based on Consortium Blockchain. In Proceedings of the Advances in Artificial Intelligence and Security, Cham, Switzerland, 19–23 July 2021; pp. 172–187.
10. Mansouri, Y.; Toosi, A.N.; Buyya, R. Data Storage Management in Cloud Environments: Taxonomy, Survey, and Future Directions. *ACM Comput. Surv.* **2017**, *50*, 91. [CrossRef]
11. Popovic, K.; Hocenski, Z. *Cloud Computing Security Issues and Challenges*; IEEE: Piscataway, NJ, USA, 2010; pp. 344–349.
12. Ren, K.; Wang, C.; Wang, Q. Security Challenges for the Public Cloud. *EEE Internet Comput.* **2012**, *16*, 69–73. [CrossRef]
13. Kotha, S.K.; Rani, M.S.; Subedi, B.; Chunduru, A.; Karrothu, A.; Neupane, B.; Sathishkumar, V.E. A Comprehensive Review on Secure Data Sharing in Cloud Environment. *Wirel. Pers. Commun.* **2022**, *127*, 2161–2188. [CrossRef]
14. Song, C.; Park, Y.; Gao, J.; Nanduri, S.K.; Zegers, W. Favored Encryption Techniques for Cloud Storage. In Proceedings of the 2015 IEEE First International Conference on Big Data Computing Service and Applications, Redwood City, CA, USA, 30 March–2 April 2015; pp. 267–274.
15. Gai, K.; She, Y.; Zhu, L.; Choo, K.-K.R.; Wan, Z. A Blockchain-Based Access Control Scheme for Zero Trust Cross—Organizational Data Sharing. *ACM Trans. Internet Technol.* **2023**, *23*, 38. [CrossRef]
16. Xu, H.; He, Q.; Li, X.; Jiang, B.; Qin, K. BDSS-FA: A Blockchain-Based Data Security Sharing Platform With Fine-Grained Access Control. *IEEE Access* **2020**, *8*, 87552–87561. [CrossRef]
17. Yang, Y.; Zhu, H.; Lu, H.; Weng, J.; Zhang, Y.; Choo, K.-K.R. Cloud based data sharing with fine-grained proxy re-encryption. *Pervasive Mob. Comput.* **2016**, *28*, 122–134. [CrossRef]
18. Qin, Z.; Xiong, H.; Wu, S.; Batamuliza, J. A Survey of Proxy Re-Encryption for Secure Data Sharing in Cloud Computing. *IEEE Trans. Serv. Comput.* **2016**, 1. [CrossRef]
19. Zhu, L.; Song, S.; Peng, S.; Wang, W.; Hu, S.; Lan, W. The Blockchain and Homomorphic Encryption Data Sharing Method in Privacy-Preserving Computing. In Proceedings of the 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD), Danang, Vietnam, 4–6 August 2022; pp. 84–87.
20. Li, R.; Song, T.; Mei, B.; Li, H.; Cheng, X.; Sun, L. Blockchain for Large-Scale Internet of Things Data Storage and Protection. *IEEE Trans. Serv. Comput.* **2019**, *12*, 762–771. [CrossRef]
21. Salim, A.; Tiwari, R.K.; Tripathi, S. An Efficient Public Auditing Scheme for Cloud Storage with Secure Access Control and Resistance Against DOS Attack by Iniquitous TPA. *Wirel. Pers. Commun.* **2021**, *117*, 2929–2954. [CrossRef]
22. Li, B.; Wu, H.; He, X.; Wang, B.; Xu, E. Survey of Storage Scalability in Blockchain Systems. *Comput. Sci.* **2023**, *50*, 318–333.
23. Zhu, L.; Gao, F.; Shen, M.; Li, Y.; Zheng, B.; Mao, H.; Wu, Z. Survey on privacy preserving techniques for blockchain technology. *J. Comput. Res. Dev.* **2017**, *54*, 2170–2186.
24. Hammoud, O.; Tarkhanov, I.A. A Novel Blockchain-Integrated Distributed Data Storage Model with Built-in Load Balancing. In Proceedings of the 2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT), Washington, DC, USA, 12–14 October 2022; pp. 1–6.
25. Asante, G.; Ben, J.; Asante, M.; Dagadu, J. A Symmetric, Probabilistic, Non-Circuit Based Fully Homomorphic Encryption Scheme. *Int. J. Comput. Netw. Appl.* **2022**, *9*, 160–168. [CrossRef]
26. Paillier, P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. In Proceedings of the Advances in Cryptology—EUROCRYPT '99, Berlin, Heidelberg, 2–6 May 1999; pp. 223–238.
27. *ISO/IEC 18033-6:2019*; IT Security Techniques—Encryption Algorithms—Part 6: Homomorphic Encryption. ISO: Geneva, Switzerland, 2019.

28. Martins, P.; Sousa, L.; Mariano, A. A Survey on Fully Homomorphic Encryption: An Engineering Perspective. *ACM Comput. Surv.* **2017**, *50*, 83. [CrossRef]
29. Khedr, A.; Gulak, G. SecureMed: Secure Medical Computation Using GPU-Accelerated Homomorphic Encryption Scheme. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 597–606. [CrossRef]
30. Blaze, M.; Bleumer, G.; Strauss, M. Divertible protocols and atomic proxy cryptography. In Proceedings of the Advances in Cryptology—EUROCRYPT'98, Berlin, Heidelberg, 31 May–4 June 1998; pp. 127–144.
31. Dalenius, T. Towards a methodology for statistical disclosure control. *Stat. Tidskr.* **1977**, *15*, 429–444.
32. Spruill, N.L. The Confidentiality and Analytic Usefulness of Masked Business Microdata. 2002. Available online: http://www.asasrms.org/Proceedings/papers/1983_114.pdf (accessed on 3 May 2024).
33. Dwork, C. Differential Privacy. In *Encyclopedia of Cryptography and Security*; van Tilborg, H.C.A., Jajodia, S., Eds.; Springer US: Boston, MA, USA, 2011; pp. 338–340.
34. Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System. *Bitcoin* **2008**, *4*, 15. Available online: https://bitcoin.org/bitcoin.pdf (accessed on 3 May 2024).
35. Stoica, I.; Morris, R.; Karger, D.; Kaashoek, M.F.; Balakrishnan, H. Chord: A scalable peer-to-peer lookup service for internet applications. *SIGCOMM Comput. Commun. Rev.* **2001**, *31*, 149–160. [CrossRef]
36. Ratnasamy, S.; Francis, P.; Handley, M.; Karp, R.; Shenker, S. A scalable content-addressable network. *SIGCOMM Comput. Commun. Rev.* **2001**, *31*, 161–172. [CrossRef]
37. Rowstron, A.; Druschel, P. Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. In Proceedings of the Middleware 2001, Berlin, Heidelberg, 12–16 November 2001; pp. 329–350.
38. Zhao, B.Y.; Kubiatowicz, J.D.; Joseph, A.D. *Tapestry: An Infrastructure for Fault-tolerant Wide-Area Location and Routing*; University of California at Berkeley: Berkeley, CA, USA, 2001.
39. Raj, T.F.M.; Vallathan, G.; Perumal, E.; Sudhakar, P.A.J. Future and Research Perspectives of Spatiotemporal Data Management Methods. In *Spatiotemporal Data Analytics and Modeling: Techniques and Applications*; A, J., Abimannan, S., El-Alfy, E.S.M., Chang, Y.S., Eds.; Springer Nature Singapore: Singapore, 2024; pp. 235–245.
40. Samanthula, B.K.; Howser, G.; Elmehdwi, Y.; Madria, S. An efficient and secure data sharing framework using homomorphic encryption in the cloud. In Proceedings of the 1st International Workshop on Cloud Intelligence, Istanbul, Turkey, 31 August 2012. [CrossRef]

information

Article

Protecting Infinite Data Streams from Wearable Devices with Local Differential Privacy Techniques

Feng Zhao and Song Fan *

Cyberspace Security Institute, Chang'an Campus, Xi'an University of Posts and Telecommunication, Xi'an 710061, China
* Correspondence: songfan@stu.xupt.edu.cn

Abstract: The real-time data collected by wearable devices enables personalized health management and supports public health monitoring. However, sharing these data with third-party organizations introduces significant privacy risks. As a result, protecting and securely sharing wearable device data has become a critical concern. This paper proposes a local differential privacy-preserving algorithm designed for continuous data streams generated by wearable devices. Initially, the data stream is sampled at key points to avoid prematurely exhausting the privacy budget. Then, an adaptive allocation of the privacy budget at these points enhances privacy protection for sensitive data. Additionally, the optimized square wave (SW) mechanism introduces perturbations to the sampled points. Afterward, the Kalman filter algorithm is applied to maintain data flow patterns and reduce prediction errors. Experimental validation using two real datasets demonstrates that, under comparable conditions, this approach provides higher data availability than existing privacy protection methods for continuous data streams.

Keywords: local differential privacy; infinite data streams; privacy protection

1. Introduction

With the rapid advancement of information technology, wearable devices have become essential tools for health monitoring, gaining global attention and widespread adoption [1]. Their real-time capabilities, affordability, and portability provide innovative solutions for both personal and professional healthcare. According to the latest Internet Data Center (IDC) data, the global and Chinese markets are experiencing continuous growth in wearable device shipments, including general health monitors, like smartwatches, and specialized medical devices, such as blood glucose meters and blood pressure monitors. These devices enable real-time health data monitoring and analysis in daily life, support the ongoing management of chronic diseases, and assist medical professionals in clinical decision-making. However, as the popularity and use of wearable medical devices continue to rise, ensuring privacy protection and data security has become increasingly critical. Users' health data, including metrics such as heart rate, physical activity levels, and sleep patterns, are not only crucial for personal health management but also contain sensitive information that requires robust protection and lawful handling.

The real-time data streams collected by wearable devices need to be aggregated and analyzed by trusted third-party organizations to fully realize their social and personal benefits [2]. Consequently, there has been a growing focus on enhancing the availability of these data while simultaneously safeguarding the privacy of wearable device users. To address this challenge, Differential Privacy (DP) [3] has been widely adopted as a robust privacy protection framework. However, DP assumes that the server is trustworthy. In practice, servers may inadvertently or deliberately compromise user privacy due to curiosity or commercial interests, leading to potential privacy breaches. To mitigate the risk of privacy leakage by the server, Local Differential Privacy (LDP) [4] was introduced. LDP

has the advantage of locally protecting a significant amount of end-user data. LDP shows great promise in enabling statistical analysis of data streams without relying on trusted third-party entities. Its goal is to preserve the privacy of individual data during both data collection and transmission. The core principle of LDP involves locally injecting noise at the data source to obscure the true values of individual data, allowing external entities to perform aggregated statistical analyses without compromising individual privacy.

However, the current method for protecting privacy in wearable device data streams using Local Differential Privacy (LDP) relies on the random response technique [5] to introduce noise into the data, thereby ensuring user privacy. While effective in protecting privacy, this approach significantly compromises data availability. To address these challenges, the authors [6] proposed the Pattern-LDP algorithm to enhance privacy protection in data streams. The algorithm first employs piecewise linear approximation to normalize the data stream and selects the furthest point within a fixed error threshold as the sampling point. However, Pattern-LDP requires continuous normalization, making it less suitable for the real-time, dynamic data streams generated by wearable devices. Additionally, when allocating the privacy budget, the algorithm only considers the speed of data stream fluctuations, neglecting the directional trends in these fluctuations, which are crucial indicators for describing dynamic real-time data streams. Furthermore, the algorithm does not optimize non-sampling data points, leading to poor availability of the published data streams and potential privacy leaks. Thus, there is an urgent need for a new and effective privacy protection method for streaming data generated by wearable devices—one that accommodates the real-time nature of dynamic data streams and allows wearable device users to locally perturb data before real-time uploading.

In response to the challenges mentioned above, this paper introduces a lightweight WIDS-LDP algorithm specifically designed for wearable devices. The algorithm consists of two primary components: the wearable device side and the device service provider side. First, on the wearable device side, significant data points are sampled based on trends and rates of fluctuation within the data stream. The privacy budget is then adaptively allocated to these sampled points, which are subsequently perturbed to protect user privacy. Second, on the device service provider side, post-processing optimization is applied to both the non-sampled and sampled points. This optimization enhances the overall utility of the published data, thereby minimizing privacy risks. The main contributions of this paper are summarized as follows:

(1) This paper proposes a local differential privacy protection framework specifically designed for wearable devices, aiming to enhance data availability while safeguarding the privacy of wearable device users.
(2) The adaptive privacy budget mechanism is optimized based on the characteristics of sampling points in the data stream, resulting in a more reasonable allocation of the privacy budget. Additionally, an improved SW mechanism is applied for perturbation, ensuring that data with smaller errors are output with higher probability.
(3) Comparisons with existing methods, using real datasets, demonstrate that this approach not only effectively protects user privacy but also preserves the availability of data streams.

2. Related Work

In this paper, we focus on local differential privacy (LDP) methods for protecting data streams. We start by introducing privacy protection techniques for sequential data streams and then provide an overview of LDP-based methods.

Privacy Protection for Time Series Data: Data stream privacy protection can be categorized into two types based on different usage scenarios: privacy protection for aggregate statistical analysis and privacy protection for time series analysis. For the first type, the authors in [7] proposed a variant of smooth projection hashing to construct a privacy protection scheme for aggregate statistical analysis. However, this scheme is relatively complex and unsuitable for wearable devices. For the second type, Zheng et al. [8] proposed a

scheme involving querying within a similar range and then reporting the results. This solution is designed for similarity queries over time series data and is not applicable to the privacy protection of sensitive data. In [9], the authors introduced a novel method that combines cryptographic algorithms with emerging data mining technologies to ensure the privacy protection of time series data. The key idea of this scheme is to utilize the observation of plaintext DTW scores and promote scalable computation in the ciphertext domain through a customized security design. However, this work focuses on the privacy protection of sensitive data, which does not align with the requirement for differential privacy. These algorithms effectively guarantee the privacy and security of time series data streams but do not consider differential privacy.

Data Stream DP: Differential Privacy (DP) was first introduced by Dwork et al. [3] to balance data availability with privacy protection requirements. Local Differential Privacy (LDP) [4] can mitigate the risk of privacy leakage associated with DP, particularly when dealing with untrustworthy third-party servers. Guan et al. [10] introduced the EDPDCS clustering scheme, which incorporates a privacy-preserving clustering method within the Map-Reduce framework. This approach uses K-means clustering combined with differential privacy (Laplace noise) to enhance the accuracy of published data. Han et al. [11] proposed the PPM-HAD algorithm, which supports operations such as (mean, variance) addition and (minimum/maximum, median) aggregation. This mechanism is particularly effective for cloud servers and can strongly resist differential attacks, but it is not applicable to wearable devices. Saleheen [12] proposed the mSieve algorithm, which integrates data-driven technology with Laplace noise to obfuscate sensitive data on demand while maintaining differential privacy. However, this approach is associated with significant error and is, therefore, not well-suited for protecting sensitive data in wearable devices. Additionally, other technologies, such as the exponential mechanism [13], Fourier algorithm [14], and classification trees [15], can be integrated with differential privacy.

Wearable Devices LDP: Kim et al. [16] proposed a privacy-preserving aggregation algorithm based on LDP. This algorithm identifies key points from the original data, adaptively adds random noise to these points, and linearly connects the noisy key point values to reconstruct data curves. Although Kim et al.'s algorithm can reconstruct data flows, it suffers from significant errors due to excessive noise and an unreasonable data curve reconstruction method. Li et al. [17] improved upon this algorithm by integrating the concept of random response with LDP and adaptively adjusting the noise magnitude based on the characteristics of the original data. However, Li's algorithm still uses a linear reconstruction method for predicting non-sampled points. Despite proposed enhancements using interpolation and fitting, the improvement remains limited. Additionally, Li's privacy budget allocation strategy employs a uniform allocation scheme, which does not fully address the protection needs of critical data, thereby posing a risk of important data leakage. Zhang et al. [18] proposed the RE-Dpocpor algorithm, which utilizes the Laplace noise mechanism combined with adaptive sampling, filtering, and budget allocation algorithms to publish differential privacy-protected real-time health data collected over w consecutive days. However, because this scheme relies on information from future timestamps, it is not applicable to infinite data streams. Tu et al. [2] proposed a mean data publishing algorithm for wearable devices that offers high availability for this purpose. However, the algorithm requires users to calculate the mean global sensitivity in advance. As a result, when dealing with different big data statistics, the global sensitivity must be recalculated, leading to reduced availability. Furthermore, calculating global sensitivity necessitates the use of the entire dataset, which is not practical for the continuous data streams generated by wearable devices. Although the aforementioned privacy protection measures for wearable devices can safeguard data flow, they each have limitations. There remains a gap in privacy protection for unlimited data streams from wearable devices when utilizing local differential privacy techniques.

Therefore, it is essential to further investigate and develop a local differential privacy protection method that ensures the published values are as close to the true values as

possible while effectively safeguarding user privacy. Additionally, the method should be applicable to wearable devices and capable of preserving the unique patterns of real-time dynamic data streams.

3. Basic Knowledge

In this section, we present the problem statement pertinent to our research topic and introduce the foundational theoretical concepts relevant to this paper.

3.1. Infinite Data Streams

Wearable devices collect data at fixed intervals, forming a data stream $S = (x_1, x_2, \ldots, x_n)$, where x_i denotes the data point at the i timestamp ($0 \leq i \leq n$), n denotes the length of the data stream, and S represents the wearable device. Device S allows users to customize timestamps and sampling intervals according to the requirements of different data types, facilitating the creation of data curves by connecting timestamped data points. Data collection and analysis are typically conducted over a defined period (e.g., 24 h) to generate a finite data stream. Wearable devices continuously collect data from users' wrists as long as the device is worn, resulting in what is referred to as an infinite data stream. When applying differential privacy protection to finite data streams, a thorough analysis based on the data's characteristics can enhance privacy protection effectiveness. However, for infinite data streams, the uncertainty of future data points requires predicting information for the next timestamp based on previous data, necessitating more sophisticated methods to ensure privacy protection.

3.2. Problem Statement

The data stream collected by a single wearable device is represented as a univariate discrete time series x. The aggregated set of multiple discrete time series x at discrete times n, where $0 \leq n \leq T$ and T is the length of the sequence, is denoted as S. S represents the aggregated sequence of raw data. For example, S could be the aggregated numerical sequences of heart rate data collected from 20 individuals over a period of time. The goal of this article is to publish the sanitized version S^* of the aggregated sequence S in real time. S^* is the published data that satisfies local differential privacy. After aggregation and analysis, S^* can provide valuable insights across various aspects. The algorithm's usage scenario is illustrated in Figure 1.

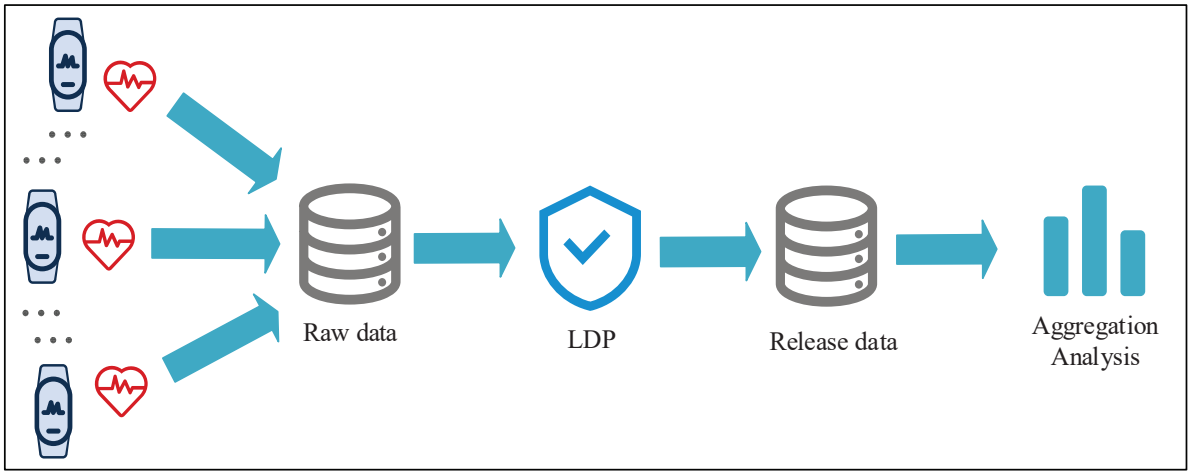

Figure 1. Usage scenario of the algorithm.

3.3. Local Differential Privacy

Local Differential Privacy (LDP) is a privacy protection mechanism focused on safeguarding the privacy of individual users or devices. Its purpose is to protect privacy during the collection and transmission of individual data. This is achieved by adding noise to the

data locally at the time of collection. The processed data, which are now privacy-protected, are then transmitted to the data collector. Typically, this provides a lighter level of privacy protection compared to other methods. The specific definition is provided in Formula (1):

$$\Pr[M(S) \in O] \leq e^{\varepsilon} \cdot \Pr[M(S') \in O] \tag{1}$$

Here, S and S' are sibling datasets that differ by at most one data point. If the probability that the result of the random algorithm M applied to these two datasets satisfies the specified formula is as described, then the random algorithm M satisfies ε-local differential privacy. Here, ε represents the privacy budget [6], which specifies the level of privacy protection provided. A smaller ε indicates a stronger privacy guarantee but introduces more noise and reduces accuracy. Conversely, a larger ε provides weaker privacy protection but allows for higher accuracy.

3.4. W-Event Privacy

W-event privacy [19]: A mechanism M is said to satisfy w-event privacy if, for any two datasets D and D' that differ in their data values on w events, and for any possible output S satisfying Formula (2), the following condition holds:

$$\Pr[M(D) \in O] \leq e^{\varepsilon} \cdot \Pr[M(D') \in O] \tag{2}$$

where ε is the privacy parameter. Here, ε quantifies the difference in the distribution of query results between D and D', and $Pr[M(D) \in O]$ represents the probability that the mechanism's output falls in the set O when the dataset is D.

4. Recommend Method

In this section, we propose a new privacy protection method for infinite data streams collected by wearable devices in real time, called WIDS-LDP (Wearables Infinite Data Stream-Local Differential Privacy).

4.1. WIDS-LDP

The framework design of WIDS-LDP is shown in Figure 2. The method includes two parts: the wearable device side and the device service provider side. First, the wearable device side performs salient point sampling, privacy budget allocation, and data perturbation. Second, the device service provider side performs post-processing optimization.

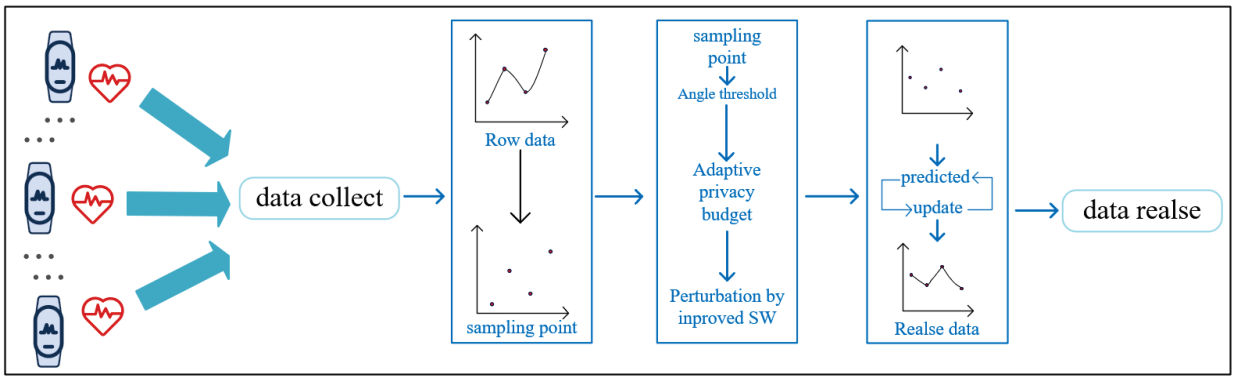

Figure 2. The framework design of WIDS-LDP.

Salient point sampling: We use a method that combines linear fitting equations with the least squares approach (hereinafter referred to as LFLS) to sample salient points, effectively representing the patterns in real-time streaming data.

Privacy budget allocation and perturbation: Based on the salient points, as well as the volatility and fluctuation amplitude of the streaming data identified in the first

part, the privacy budget is adaptively allocated, and perturbation is applied using the SW mechanism.

Post-processing optimization: Kalman filtering is used to predict non-sampling point data, while the perturbed sampling point data are predicted and updated to enhance accuracy.

4.2. Significant Point Sampling

In this section, we provide a detailed introduction to the LFLS algorithm. The algorithm begins by using the least squares method [20] to determine the mathematical representation of the linear fitting equation for the current significant point. Next, based on this linear fitting equation, it assesses whether the next significant point can be fitted into the current equation. This approach effectively models the fluctuation pattern of the flow data, allowing for accurate sampling of the pattern's significant points.

4.2.1. First-Order Difference Method

Assuming the data stream is $S = \{x_1, x_2, \ldots, x_n\}$, this method can determine whether the point at time t is significant once the point at time $t+1$ is known. The steps are as follows: First, calculate the first-order difference $Diff_i = x_t - x_{t-1}$ corresponding to each data point in the stream. When the sign of the slope of the linear fitting equation between x_t and x_{t+1}, and x_{t-1} changes, x_t is identified as a sampling point.

4.2.2. LFLS

The fitting process of the LFLS algorithm is as follows: First, we select two adjacent significant points $A(a, x_a)$ and $B(b, x_b)$, and derive the fitting straight line equation $F(x) = kx + b$, where the slope k is calculated as $k = (x_b - x_a)/(b - a)$. For the next significant point $C(c, x_c)$, the straight-line equation connecting B and C is given by $F(x') = kx' + b$, where $k = (x_c - x_b)/(c - b)$. Assume that the angle between $F(x)$ and $F(x')$ does not exceed a certain threshold, meaning

$$\tan\theta = \left| \frac{k - k'}{1 + kk'} \right| \leq \tan\alpha \tag{3}$$

If the angle between the fitting equation of A and B and the line connecting B and C does not exceed the specified threshold, C can be considered to fit the linear equation of A and B. This allows us to calculate the best-fitting line for points A, B, and C. However, if the angle between the fitting lines of A and B and C exceeds the threshold, the linear fitting equation is recalculated starting from C. By applying this process, we generate a set of linear fitting equations. If adjacent fitting lines exhibit the same trend, the point is not deemed significant.

4.2.3. Dynamic Angle α

We use a PID controller to represent the fluctuation rate of the data stream. A greater fluctuation rate indicates a faster rate of change of the data flow, while a smaller fluctuation rate reflects a slower rate of change. The complete PID algorithm is shown below:

$$\Delta_{k_n} = C_p E_{k_n} + \frac{C_i}{T_i} \sum_{j=n-T_i+1}^{n} E_{k_j} + C_d \frac{E_{k_n} - E_{k_{n-1}}}{k_n - k_{n-1}} \tag{4}$$

Among them, C_p represents the gap between the target value and the actual value, C_i represents the accumulation of error over time, and C_d represents the error in predicting future values. T_i represents the number of errors in the cumulative integral error. E_{k_i} is the feedback error, expressed as $E_{k_i} = |x_t - x'_t|$, t represents the timestamp, x'_t represents the predicted value.

The formula $\alpha = \lambda\pi/2$ indicates that changes in α lead to corresponding changes in λ. To dynamically adjust λ, we can change α. Since λ needs to remain within the range

(0,1), we use an exponential function to determine its value. Accordingly, λ can be defined as follows:

$$\lambda = 1 - \exp(-(\frac{1}{|k|} + \Delta_{k_n})) \tag{5}$$

Therefore, depending on the trend and rate of data flow, we can dynamically change the angle threshold to adaptively change the sampling interval to balance privacy and utility. The specific salient point sampling is shown in Algorithm 1.

Algorithm 1: Significant point sampling

Input: Raw data $S = (x_1, x_2, \ldots, x_n)$,$\alpha$
Output: S'
for t ∈ [2,n] do
  if $t + 1 < n$ then
    $k = x_t - x_{t-1}$, $k' = x_t - x_{t-1}$
  end
  Calculate $\tan\theta = \left|\frac{k-k'}{1+kk'}\right|$
  if tan θ < tan α
    $x_t \in S$
  end
end

4.3. Budget Allocation and Perturbations

4.3.1. Adaptive Privacy Budget Allocation

We use the LBD (LDP Budget Distribution) model to adaptively allocate privacy budgets based on the characteristics of the sampling points. For a single sliding window, the privacy budget ε is evenly distributed to the difference budget and the release budget. First, the difference budget is evenly distributed to each timestamp. The perturbed data are used to predict the difference error. Then, the remaining privacy budget for the current timestamp is calculated and the difference budget is adaptively allocated. The specific adaptive privacy budget allocation is shown in Algorithm 2.

Algorithm 2: Adaptive privacy budget allocation

Input: Privacy budget ε, window size ω
Output: Perturbation data $p = (p_1, p_2, \ldots, p_n)$
Initialize $i = 1$
for $t \in [t-1, n]$ do
  $\varepsilon_{t,1} = \varepsilon/(2\omega)$
  Calculate remaining publication budget $\varepsilon_{rm} = \frac{\varepsilon}{2} - \sum_{i=t-\omega+1}^{t-1} \varepsilon_{t,2}$
  $\varepsilon_{t,2} = \varepsilon_{rm}/2$
  $\varepsilon_i = \varepsilon_{t,1} + \varepsilon_{t,2}$
end

4.3.2. Data Perturbations

Gao et al. [21] applied an enhanced version of the SW mechanism [22] to perturb data streams. This improved method updates the perturbation probability and range of the traditional SW mechanism, resulting in perturbed data that more closely align with the original data curve. The mechanism is defined as follows:

$$b_i = (\varepsilon_i e^{\varepsilon_i} - e^{\varepsilon_i} + 1)/(2e^{\varepsilon_i}(e^{\varepsilon_i} - 1 - \varepsilon_i)) \tag{6}$$

The disturbance probability is expressed in Equation (7).

$$\begin{cases} p_i = e^{\varepsilon_i}/2b_i e^{\varepsilon_i} + 1, \text{ if } |x_i - \widetilde{x}_i| \le b_i \\ q_i = (2b_i e^{\varepsilon_i} + 1)^{-1}, \textit{ otherwise} \end{cases} \tag{7}$$

Therefore, the original data with smaller prediction errors are output with probability p_i, while the original data with larger prediction errors are output with probability q_i. This approach not only minimizes the introduction of noise but also ensures user privacy

4.4. Post-Processing Mechanism

Kalman et al. [23] proposed a method to solve linear filtering and prediction problems using linear state equations, called Kalman filtering. The algorithm consists of two parts: prediction and update.

Prediction: Estimate the state at the current time based on the posterior estimate (update value) at the previous time and derive the prior estimate (prediction value) at the current time. The specific process is as follows:

$$\widetilde{x}_t = A\hat{x}_{t-1} + Bu_{t-1} \tag{8}$$

$$\widetilde{R}_t = A\hat{P}_{t-1}A^T + Q \tag{9}$$

Here, $\hat{x}_{t-1}$ is the filtering result, also called the best result. A represents the state transfer matrix, and B represents the input control matrix. $\hat{x}_{t-1}$ represents the external operation at timestamp $t-1$. Q is the covariance of the state transition. Updating means using the measured value at a certain moment to correct the predicted system state. The specific process is as follows:

$$K_t = \frac{\widetilde{P}_t H^T}{H\widetilde{P}H^T + R} \tag{10}$$

$$\hat{x}_t = \hat{x}_t + K_t(\overline{x}_k - H\widetilde{x}_t) \tag{11}$$

$$\hat{P}_t = (1 - K_t H)\widetilde{P}_t \tag{12}$$

Here, H represents the transformation matrix of the state variables. K_t represents the Kalman gain. Algorithm 3 is shown below.

Algorithm 3: Adaptive privacy budget allocation

Input: Perturbation data $p = (p_1, p_2, \ldots, p_n)$
Output: Release data $r = (r_1, r_2, \ldots, r_n)$
for $t \in [t-1, n]$ do
$\hat{x}_i^- = \hat{x}_{i-1}$
$P_i^- = p_{i-1} + Q$
$K_K = P_i^-\left(P_i^- + R\right)^{-1}$
$\hat{x}_i = \hat{x}_i^- + K_K\left(x_i - \hat{x}_i^-\right)$
$P_i = \left(1 - K_K\right) P_i^-$
$r_i = \hat{x}_i$
end

4.5. Theoretical Analysis

Theorem 1. *Serial Combination Theorem.*

Proof. In the dataset D, assuming that the algorithm contains M random algorithms A_i, A_i satisfies ε_i-differential privacy, and the random processes between mechanisms are independent of each other, then the algorithm satisfies $\sum_{1 \le i \le M} \varepsilon_i$-differential privacy. □

Theorem 2. *WIDS-LDP satisfies ε-LDP.*

Proof. In the WIDS-LDP algorithm, the perturbation module and the privacy budget allocation module access the original data, and the others are operations on the perturbed data. According to [24], as long as the post-processing algorithm does not directly use the original data information, the post-processing algorithm is privacy-preserving. Therefore,

if we can prove that the perturbation and privacy budget allocation modules satisfy ε-local differential privacy, then the solution in this paper satisfies it. □

According to [21], the SW perturbation module satisfies ε_i-local differential privacy. According to Algorithm 2 and Theorem 1, the whole algorithm satisfies $\sum_{1\leq i\leq n}\varepsilon_i$-local differential privacy. Also, because $\sum_{1\leq i\leq n}\varepsilon_i \leq \varepsilon$, the algorithm satisfies ε-LDP.

5. Experiment

In this section, we first describe the specific experimental settings of this study, followed by an introduction to the comparative scheme. Finally, we evaluate the performance of the proposed scheme using two real datasets. The evaluation focuses on three key aspects: (1) the impact of different window sizes on the error rate; (2) the impact of different privacy budgets on the error rate; and (3) the impact of different data lengths on the error rate. These results will provide a crucial basis for understanding the effectiveness and applicability of the proposed scheme.

5.1. Experimental Environment

The experimental part of this paper is completed on a personal computer equipped with an Intel(R) Core(TM) i7-8565UCPU, 8 GB RAM, and a 64-bit Windows 11 operating system. The algorithm is implemented using MATLAB 2020a and is compiled and run in this environment.

5.2. Real Dataset

PAMAP [25]: The PAMAP dataset has 9 subjects (8 males and 1 female) wearing three inertial measurement units and heart rate monitors to record 18 activities, with a total of more than 10 h of data collected. The heart rate data of 8 people were selected from the dataset as the raw data streams of the experiment, and the length of each data stream was 3000 (Table 1). The test subjects were numbered from 1 to 8 and recorded once every minute, so the total data were 3×8 K.

Table 1. PAMAP data range table.

Tester	1	2	3	4	5	6	7	8
Size	3000	3000	3000	3000	3000	3000	3000	3000
Range	78~120	74~107	68~94	57~121	70~101	60~104	60~99	66~104

Taxi [26]: The dataset contains the real-time movement trajectories of 10,357 taxis. The real-time location was extracted every 10 min, with a total of 886 timestamps. The area was divided into 5 grids, that is, T = 5, and we obtained d = 10,357 data streams for each taxi.

Heart rate data are a crucial indicator for wearable devices in monitoring users' health status, as they reflect physiological conditions, activity levels, and their changes. Additionally, the continuous nature of the data collection process aligns with the definition of a data stream. The PAMAP dataset has been extensively utilized in numerous related studies, and its findings are well-recognized, allowing us to compare our research with existing results to validate the effectiveness and innovation of our proposed method. Similarly, the TAXI dataset provides location information, and its data collection process also adheres to the definition of a data stream, which is pertinent to the research direction of this article.

5.3. Comparison Scheme

In this section, this article not only compares WIDS-LDP and PP-LDP, but also the following two solutions:

LDP Budget Distribution (LBD) [27]: The scheme allocates the privacy budget in an exponentially decreasing manner. The perturbation data distribution is $h = (d - 2 + e^{\varepsilon})/\left((e^{\varepsilon} - 1)^2\right)$, $p_i \in [x_i - h, x_i + h]$.

Piecewise Mechanism (PM) [28]: The data perturbation of this scheme is defined as $h = \left(\frac{4}{3}e^{\frac{\varepsilon}{2}}/\left(e^{\frac{\varepsilon}{2}} - 1\right)^2\right)$, $p_i \in [x_i - h, x_i + h]$.

5.4. Experiment Indicatorsr

This paper selects the mean relative error (MRE) as an indicator to measure the experimental error.

$$\mathrm{MRE} = \frac{1}{\mathrm{n}} \times \sum_{\mathrm{d}=1}^{\mathrm{n}} \frac{|\,\mathrm{AVG}_{\mathrm{actual}}(\mathrm{x_d}) - \mathrm{AVG}_{\mathrm{est}}(\mathrm{x_d})|}{\mathrm{AVG}_{\mathrm{actual}}(\mathrm{x_d})} \tag{13}$$

AVGest(x_d) and AVGactual(x_d) represent the estimated average and the actual average of x_d at timestamp t_d, respectively, and n denotes the sequence length. MRE is used as an indicator to measure data availability, with smaller values indicating lower errors and higher availability.

5.5. Experiment Parameters

In the adaptive sampling stage, this paper uses a PID controller to sample significant points according to the data flow fluctuation trend. We set the PID parameters as follows: C_P = 0.8, C_i = 0.1, and C_d = 0.1. When evaluating the solution of this article, this article uses the control variable method to evaluate the effectiveness of this method in generating real-time data streams on wearable devices for different privacy budgets, different sliding window lengths, and different data stream lengths. Each experiment was run 100 times, and the results were averaged. The privacy budget (defaults to 1 in other cases) and the sliding window length (defaults to 20 in other cases) in this article replicate the dataset to simulate different data stream lengths (Table 2).

Table 2. Parameter setting table.

Parameter	C_P	C_i	C_d	ε	ω
Range	0.8	0.1	0.1	[0.5, 2.5]	[10, 50]
Default	0.8	0.1	0.1	1	20

5.6. Program Utility

5.6.1. MRE Analysis of Different Window Lengths

As shown in Figure 3, the impact of different window lengths on MRE is illustrated. Four different schemes were tested on the PAMAP and Taxi datasets with a privacy budget of 1 and data length of 20×10^4. As the sliding window length increases, the MRE gradually increases. This is because a longer sliding window contains more sampling points, leading to a smaller privacy budget allocated per point, which, in turn, increases the error. Therefore, the sliding window length is directly proportional to the MRE.

The PM and LBD algorithms exhibit larger errors compared to WIDS-LDP and PP-LDP across all window lengths. This is because WIDS-LDP and PP-LDP employ post-processing algorithms for optimization after perturbation, allowing them to predict and correct the perturbed data, resulting in lower errors. In the PAMAP dataset experiment, as shown in the left figure of Figure 3, the PM and LBD algorithms display significant fluctuations when the window length is between [10, 30], indicating that these algorithms are less stable across different datasets. In contrast, WIDS-LDP and PP-LDP demonstrate greater stability and are better suited for enhancing privacy protection and data availability.

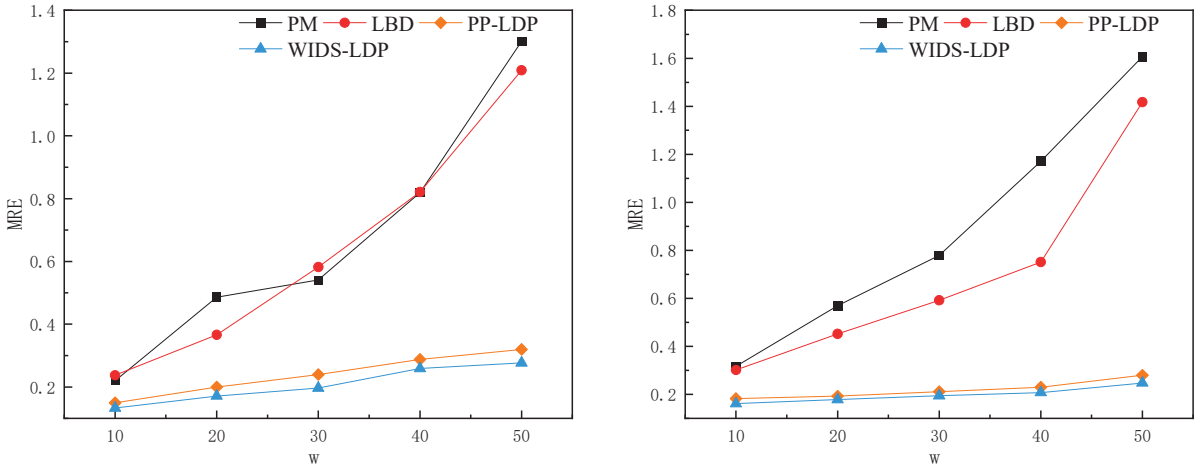

Figure 3. The impact of different window lengths on MRE (**left**: PAMAP, **right**: Taxi).

5.6.2. MRE Analysis of Different Privacy Budgets

As shown in Figure 4, the impact of different privacy budgets on MRE is illustrated. Four different schemes were tested on various datasets with a sliding window length of 20 and data length of 20×10^4. As the privacy budget increases, the MRE decreases. This is because a higher privacy budget per window allows for a larger allocation of the budget to each sampling point, reducing the error, provided other conditions remain unchanged. Thus, the privacy budget is inversely proportional to the MRE.

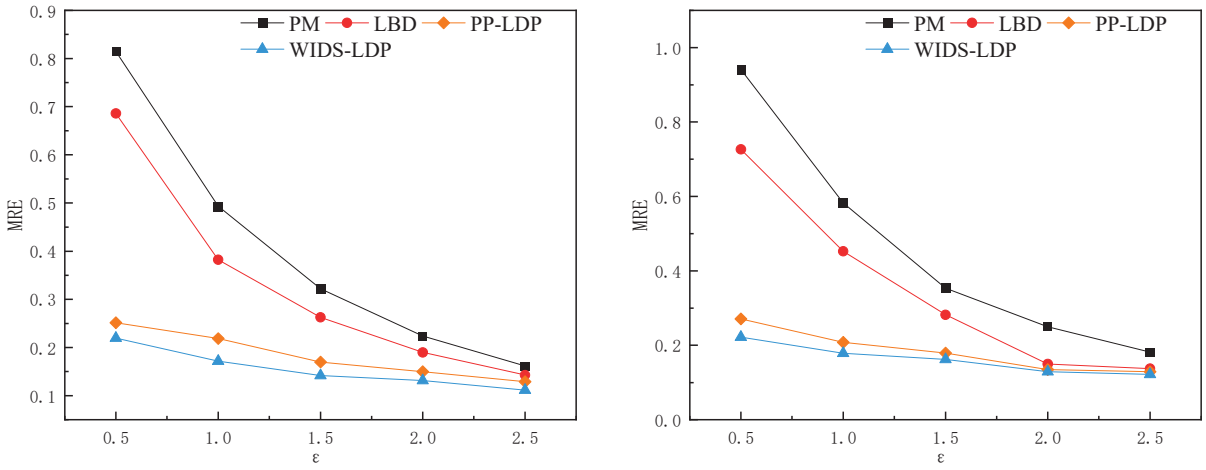

Figure 4. The impact of different privacy budgets on MRE (**left**: PAMAP, **right**: Taxi).

In these experiments, the PM and LBD algorithms show the highest error when the privacy budget is 0.5. As the privacy budget increases, the error decreases, reaching its lowest point at a budget of 2.5. However, a very high privacy budget is not ideal, as it compromises privacy protection, which must be avoided. In the Taxi dataset, as illustrated in the right figure of Figure 4, the LBD algorithm experiences significant fluctuations when the privacy budget is 2. This fluctuation is similar to the errors observed with the PP-LDP and WIDS-LDP algorithms, suggesting that while privacy protection is effective, the error with LBD is greater compared to PP-LDP and WIDS-LDP in other scenarios. WIDS-LDP proves to be more suitable for wearable device environments after optimizing the privacy budget solution.

5.6.3. MRE Analysis of Different Data Lengths

As shown in Figure 5, this paper replicates the experimental dataset to simulate an infinite data flow scenario. Four different schemes were tested on various datasets with a sliding window length of 20 and a privacy budget of 1. As the number of data flows increases, the MRE gradually decreases. This is because, with longer data streams, previous data provides insights for processing subsequent data, resulting in reduced error over time. This characteristic aligns with the privacy protection requirements for infinite data streams from wearable devices, where the data characteristics of users are generally stable.

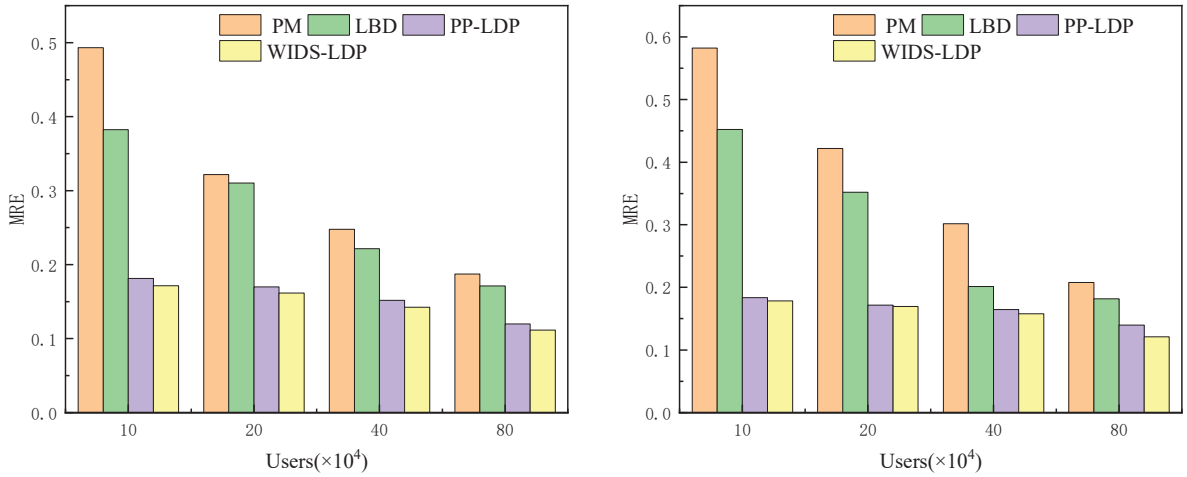

Figure 5. The impact of different data lengths on MRE (**left**: PAMAP, **right**: Taxi).

In these experiments, the PP-LDP and WIDS-LDP algorithms consistently show lower errors compared to PM and LBD across different data lengths. Nevertheless, as the data stream increases, the error decreases for all algorithms, indicating that all methods benefit from leveraging previous data experience.

In summary, WIDS-LDP and PP-LDP demonstrate stability across different experimental setups, indicating that both schemes are relatively robust and well suited for the privacy protection of infinite data streams. The WIDS-LDP scheme achieves a lower MRE compared to PP-LDP due to its optimized privacy budget allocation mechanism. Consequently, WIDS-LDP offers better data availability while ensuring privacy compared to existing schemes. For real-time dynamic infinite data streams from wearable devices, the WIDS-LDP scheme manages larger data volumes with progressively smaller MRE and higher data availability over time. The advantages and disadvantages of the proposed solutions are shown in Table 3.

Table 3. Solutions comparison.

Solution	Advantage	Disadvantage
PM	Supports multi-value and multi-attribute data	Not applicable for wearable devices; Unable to maintain the data flow pattern
LBD	Population division-based and data-adaptive algorithms; Two privacy budget allocation methods are more reasonable	Unable to maintain the dataflow pattern; The perturbation scheme has large errors; Not applicable for wearable devices
PP-LDP	Optimized SW data perturbation method; Maintain data flow patterns	Exponentially decreasing privacy budget allocation method
WIDS-LDP	A framework suitable for wearable devices; Maintain data flow patterns; LBD privacy budget allocation method	Only supports single-dimensional data

6. Discussion

This paper primarily investigates the WIDS-LDP privacy protection method for dynamic, unlimited data streams collected by wearable devices and proposes a privacy protection framework specifically for these devices. The WIDS-LDP algorithm is implemented on both the wearable device management side and the user side, aiming to enhance data availability while safeguarding user privacy. The WIDS-LDP algorithm first identifies potential salient points using a linear fitting method. It then employs a PID controller to calculate an adaptive threshold based on previous data flows, dynamically sampling subsequent data flows and updating the threshold accordingly. This adaptive threshold ensures that data from previous points meet the criteria for a dynamic infinite data flow. Subsequently, an improved SW mechanism is used to probabilistically perturb the sampling points according to their characteristics, ensuring that user privacy is maintained. Finally, the Kalman filter mechanism is applied for post-processing optimization to prevent the inference of useful information from disturbed data points and to reduce prediction errors. Experiments conducted on two real datasets demonstrate that the WIDS-LDP scheme not only provides superior privacy protection but also enhances data availability. The WIDS-LDP framework contributes significantly to the expanding field of privacy-preserving data management by providing effective solutions tailored to dynamic data flows. Its implementation optimizes data availability while ensuring robust privacy protection, which is crucial for applications that rely on big data analytics.

The WIDS-LDP framework can be effectively applied across various domains, including health monitoring, fitness tracking, and personalized healthcare. In clinical settings, wearable devices continuously collect patient data while ensuring privacy, enabling real-time health monitoring without compromising sensitive information. In fitness applications, the WIDS-LDP framework allows users to share their performance data for analysis and improvement while maintaining personal privacy. This framework empowers users by enabling them to contribute to aggregated performance metrics without revealing their individual data, thus facilitating personalized training recommendations based on group performance trends.

Additionally, the framework offers solutions to common challenges in data security and privacy. By employing adaptive thresholding and probabilistic perturbation techniques, the WIDS-LDP algorithm preserves privacy while maintaining data utility. This approach enhances data security for users and improves data availability for analytics, allowing organizations to derive meaningful insights from large datasets without compromising individual privacy. Furthermore, implementing this framework can foster user trust, as individuals are more likely to adopt wearable technologies when they are confident that their personal data are protected.

Overall, the WIDS-LDP framework not only enhances privacy and security in data collection but also opens new avenues for data-driven decision making in health and fitness. Consequently, it encourages broader adoption of wearable technologies and supports the development of innovative health solutions that can significantly improve patient outcomes and enhance user experience.

Limitations: Despite these advances, this study has several limitations. The public dataset used for testing is relatively short and may not adequately capture the complexity of real-world data flows, potentially affecting the robustness of the results. Additionally, the algorithm may encounter errors when applied in real-world scenarios, impacting its effectiveness in certain cases. For instance, the algorithm's assumptions about data distribution may not hold true across all operating environments, which could lead to inaccurate predictions or suboptimal performance.

Future work: To address these limitations, future research should prioritize the integration of diverse and comprehensive datasets that accurately reflect the complexity of real-world data flows. This effort could involve forming collaborations with industry partners to access proprietary datasets and employing synthetic data generation techniques

to create more representative samples. By doing so, researchers can capture a broader range of variables and conditions that influence data behavior.

Furthermore, conducting rigorous cross-validation across various datasets will significantly enhance the reliability of the findings and support broader generalizations. This approach not only strengthens the robustness of the results but also ensures that the conclusions drawn are applicable to different contexts and scenarios. Ultimately, such methodological improvements will contribute to a deeper understanding of the underlying phenomena and promote more effective solutions to the challenges identified in this study.

Future work can also focus on a detailed analysis of error bounds across various application scenarios. By evaluating the performance of the WIDS-LDP algorithm in different environments, we can enhance the practical usability of data release while maintaining strong privacy protection. Such exploration is critical for adapting the framework to meet the specific needs of various applications in real-world settings.

Author Contributions: Conceptualization, S.F. and F.Z.; methodology, S.F.; software, S.F.; formal analysis, S.F.; investigation, S.F.; writing—original draft preparation, S.F.; supervision, F.Z.; project administration, F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in [Github] at https://github.com/songsmallsong/Protecting-Infinite-Data-Streams-from-Wearable-Devices-with-Local-Differential-Privacy-Techniques (accessed on 8 September 2024). The datasets used in this paper are not strictly data streams, as the data points are discrete. These datasets were selected due to limitations in experimental equipment and the need for convenient verification. Despite having fixed timestamps and intervals, the data collection process is still real-time and continuous. Consequently, these datasets reflect the dynamic, continuous, and near-real-time characteristics of data streams. The choice of these smaller datasets was driven by computing resource constraints, but the methodology is designed for broader application to larger-scale data streams.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Terminology	Abbreviation	Definition
Square wave	SW	A periodic waveform whose value rapidly switches between two fixed levels, commonly used in signal processing and test systems.
Kalman filtering	KF	An algorithm for estimating the state of a dynamic system that achieves recursive estimation of the system state by modeling measurement noise and system noise.
Mean relative error	MRE	A measure of prediction accuracy that calculates the average of the relative error between the predicted value and the actual value and is used to evaluate the performance of the model.
Linear fitting equations with least squares	LFLS	A statistical method that finds the best-fitting line by minimizing the squared difference between the observed data and the fitted model. It is widely used in data analysis and regression modeling.
LDP budget distribution	LBD	Refers to the allocation strategy of privacy budget (or noise level) in local differential privacy, aiming to balance the privacy protection and information utility of data.
Differential privacy	DP	A method of protecting personal privacy by adding random noise to query results to ensure that individual participation does not significantly affect the overall data analysis results.

Local differential privacy	LDP	An implementation of differential privacy that allows users to perturb their data locally, ensuring that the privacy of the data is protected before being transmitted to the server.

References

1. Babu, M.; Lautman, Z.; Lin, X.; Sobota, M.H.; Snyder, M.P. Wearable devices: Implications for precision medicine and the future of health care. *Annu. Rev. Med.* **2024**, *75*, 401–415. [CrossRef] [PubMed]
2. Tu, Z.X.; Liu, S.B.; Xiong, X.X. Differential privacy mean publishing of digital stream data for wearable devices. *Comput. Appl.* **2020**, *40*, 6.
3. Dwork, C.; Mcsherry, F.; Nissim, K. Calibrating noise to sensitivityn in private data analysis. In Proceedings of the Theory of Cryptography: Third Theory of Cryptography Conference, New York, NY, USA, 4–7 March 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284.
4. Kasiviswanathan, S.P.; Lee, H.K.; Nissim, K. What can we learn privately? *SIAM J. Comput.* **2011**, *40*, 793–826. [CrossRef]
5. Yan, Y.; Chen, J.; Mahmood, A.; Qian, X.; Yan, P. LDPORR: A localized location privacy protection method based on optimized random response. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101713. [CrossRef]
6. Wang, Z.; Liu, W.; Pang, X. Towards pattern-aware privacy-preserving real-time data collection. In Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications, Virtual, 6–9 July 2020; pp. 109–118.
7. Benhamouda, F.; Joye, M.; Libert, B. A new framework for privacy-preserving aggregation of time-series data. *ACM Trans. Inf. Syst. Secur. (TISSEC)* **2016**, *18*, 1–21. [CrossRef]
8. Zheng, Y.; Lu, R.; Guan, Y. Efficient and privacy-preserving similarity range query over encrypted time series data. *IEEE Trans. Dependable Secur. Comput.* **2021**, *19*, 2501–2516. [CrossRef]
9. Liu, Z.; Yi, Y. Privacy-preserving collaborative analytics on medical time series data. *IEEE Trans. Dependable Secur. Comput.* **2020**, *19*, 1687–1702. [CrossRef]
10. Guan, Z.T.; Lv, Z.F.; Du, X.J.; Wu, L.F.; Guizani, M. Achieving data utility-privacy trade off in Internet of medical things, a machine learning approach. *Future Gener. Comput. Syst.* **2019**, *98*, 60–68. [CrossRef]
11. Song, H.; Shuai, Z.; Qinghua, L. PPM-HDA: Privacy-preserving and multifunctional health data aggregation with fault tolerance. *IEEE Trans. Inf. Forensics Secur.* **2016**, *18*, 1940–1955.
12. Saleheen, N.; Chakraborty, S.; Ali, N.; Rahman, M.M.; Hossain, S.M.; Bari, R.; Buder, E.; Srivastava, M.; Kumar, S. mSieve: Differential behavioral privacy in time series of mobile sensor data. In Proceedings of the 2016 ACM International Joint Conference, Heidelberg, Germany, 12–16 September 2016; ACM: Heidelberg, Germany, 2016; pp. 706–717.
13. Steil, J.; Hagestedt, I.; Huang, M.X.; Bulling, A. Privacy aware eye tracking using differential privacy. In Proceedings of the ACM. the 11th ACM Symposium, Denver, CO, USA, 20–25 June 2019; ACM: New York, NY, USA, 2019; pp. 1–9.
14. Bozkir, E.; Günlü, O.; Fuhl, W.; Schaefer, R.F.; Kasneci, E. Differential privacy for eye tracking with temporal correlations. *PLoS ONE* **2021**, *16*, e0255979. [CrossRef] [PubMed]
15. Zhang, S.Q.; Li, X.H. Differential privacy medical data publishing method based on attribute correlation. *Sci. Rep.* **2022**, *12*, 15725. [CrossRef] [PubMed]
16. Kim, J.W.; Jang, B.; Yoo, H. Privacy-preserving aggregation of personal health data streams. *PLoS ONE* **2018**, *13*, e0207639. [CrossRef] [PubMed]
17. Li, Z.B.; Wang, B.H.; Li, J.S. Local differential privacy protection for wearable device data. *PLoS ONE* **2022**, *17*, e0272766. [CrossRef] [PubMed]
18. Zhang, J.; Liang, X.; Zhang, Z.; He, S.; Shi, Z. Re-DPoctor: Real-time health data releasing with w-day differential privacy. In Proceedings of the IEEE.GLOBECOM 2017—2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–6.
19. Schäler, C.; Hütter, T.; Schäler, M. Benchmarking the Utility of w-Event Differential Privacy Mechanisms-When Baselines Become Mighty Competitors. *Proc. VLDB Endow.* **2023**, *16*, 1830–1842. [CrossRef]
20. Ding, F. Least squares parameter estimation and multi-innovation least squares methods for linear fitting problems from noisy data. *J. Comput. Appl. Math.* **2023**, *426*, 115107. [CrossRef]
21. Gao, W.; Zhou, S. Privacy-Preserving for Dynamic Real-Time Published Data Streams Based on Local Differential Privacy. *IEEE Internet Things J.* **2023**, *11*, 13551–13562. [CrossRef]
22. Li, Z.; Wang, T.; Lopuhaä-Zwakenberg, M.; Li, N.; Škoric, B. Estimating numerical distributions under local differential privacy. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, Portland, OR, USA, 14–19 June 2020; pp. 621–635.
23. Khodarahmi, M.; Maihami, V. A review on Kalman filter models. *Arch. Comput. Methods Eng.* **2023**, *30*, 727–747. [CrossRef]
24. Shanmugarasa, Y.; Chamikara MA, P.; Paik, H.; Kanhere, S.S.; Zhu, L. Local Differential Privacy for Smart Meter Data Sharing with Energy Disaggregation. In Proceedings of the 2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Abu Dhabi, United Arab Emirates, 29 April–1 May 2024; pp. 1–10.
25. Reiss, A.; Stricker, D. Introducing new benchmarked dataset for activity monitoring. In Proceedings of the IEEE, The 16th International Symposium on Wearable Computers, ISWC 2012, Newcastle Upon Tyne, UK, 18–22 June 2012; pp. 108–109.

26. Available online: https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/ (accessed on 8 September 2024).
27. Ren, X.; Shi, L.; Yu, W. LDP-IDS: Local differential privacy for infinite data streams. In Proceedings of the 2022 International Conference on Management of Data, Philadelphia, PA, USA, 12–17 June 2022; pp. 1064–1077.
28. Wang, N.; Xiao, X.; Yang, Y.; Zhao, J.; Hui, S.C.; Shin, H.; Shin, J.; Yu, G. Collecting and analyzing multidimensional data with local differential privacy. In Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 8–11 April 2019; pp. 638–649.

information

MDPI

Article

Smartphone Privacy and Cyber Safety among Australian Adolescents: Gender Differences

Yeslam Al-Saggaf * and Julie Maclean

School of Computing, Mathematics and Engineering, Charles Sturt University, Wagga Wagga, NSW 2678, Australia; jumaclean@csu.edu.au
* Correspondence: yalsaggaf@csu.edu.au

Abstract: While existing studies explore smartphone privacy setting risks for adolescents, they provide limited insight into the role of gender in these dynamics. This study aims to enhance adolescents' awareness of the security risks associated with smartphone privacy leakage by focusing on how a cyber safety intervention lesson can affect knowledge of smartphone privacy settings, attitudes toward smartphone settings, and concerns about smartphone privacy. This study surveyed 376 high school students before and after a cyber safety lesson. Our study found that before the cyber safety intervention, females reported lower knowledge of smartphone settings than males. After the lesson, this gap narrowed, with both genders demonstrating more consistent understanding. Both genders showed lower attitudes towards smartphone privacy compared to knowledge, with males displaying the largest gap, reflecting the privacy paradox. Females expressed greater concern regarding location privacy, especially when tracked by unknown individuals, indicating that while both genders are aware of risks, females perceive them more acutely. The results suggest that targeted educational programs can effectively enhance adolescents' knowledge, attitudes, and concerns about smartphone privacy, particularly in technical areas where gender gaps exist.

Keywords: smartphones; information security management; privacy; cybersecurity; Internet of Things (IoT); adolescent; gender

1. Introduction

A recent survey revealed that 95% of U.S. teens have access to smartphones, with 94% of adolescent boys and 97% of adolescent girls reporting smartphone ownership [1]. While this widespread access enables connectivity and learning, it also introduces significant privacy risks, especially if smartphone data leakage is not properly managed. Research shows that privacy behaviour in online environments is shaped by various factors, including privacy concerns, attitudes, knowledge, skills, experience, education, gender, and age [2,3], trust and risk perceptions [4] and offline expected social norms [5]. Children primarily use their smartphones to pass the time, connect with others, and learn new things [6]. However, frequent use of apps for social media, gaming, and communication often leads to the neglect of critical privacy settings, thereby increasing the risk of personal information leakage [7,8].

Studies in child psychology reveal that children's impulsive behaviour and strong desire for social connectivity often overshadow their privacy concerns [9]. Male and female teenagers exhibit significantly different attitudes toward privacy, influenced by socialization, personal experiences, and perceived risks [10]. Despite being aware of risks such as exposure to inappropriate content, interactions with strangers, and oversharing, children often struggle to take effective steps to address privacy issues and prevent data tracking [11]. This behaviour reflects the "privacy paradox", where children express concern for privacy but engage in actions that compromise it. Therefore, effective educational interventions are crucial to raising children's awareness of privacy risks and encouraging protective behaviours.

In this study, smartphone privacy refers to the protection of sensitive personal information that smartphones collect, store, and share. It focuses on the awareness and behaviours of adolescents in managing privacy settings to prevent unauthorized access or exposure of their data, such as location, and personal identifiers. Specific privacy risks include leaving Wi-Fi and Bluetooth enabled, sharing real names for device identification, and failing to manage location tracking settings. This study examines the knowledge and behaviours of adolescents regarding these risks and how targeted educational interventions can improve their understanding and practices to better protect their privacy.

Despite the importance of safeguarding children's smartphone privacy, there is limited research on how educational interventions can effectively protect children, particularly in the context of gender differences. This study aims to enhance adolescents' awareness of the security risks associated with smartphone privacy leakage by focusing on three key areas: (1) knowledge of smartphone privacy settings, (2) attitudes toward smartphone settings, and (3) concerns about smartphone privacy. Critically, this study investigates gender differences in adolescents' responses to a hands-on cyber safety lesson they attended in their school to raise their awareness about these security areas.

Although existing studies examine adolescent risk-taking in relation to privacy [9], online privacy perceptions [10], and general privacy attitudes [11], they offer limited insight into the specific role gender plays in smartphone privacy dynamics. This study builds on the protection motivation theory (PMT) [12], which posits that individuals are motivated to protect themselves based on their perceived severity and vulnerability to a threat. This theoretical framework is particularly relevant to online privacy, as it helps identify factors that may deter individuals from engaging in protective behaviours. PMT may be instrumental in addressing gender disparities in adolescent privacy behaviours and in developing targeted strategies to reduce the risk of smartphone privacy leakage across different genders. As technology continues to evolve and new privacy threats emerge, it is imperative to continuously update educational content to effectively safeguard children's privacy.

This paper is structured into five key sections. The Section 2 reviews the existing literature on smartphone privacy, cyber safety education, and gender differences in digital literacy among adolescents. In Section 3, we describe the survey methodology, the educational interventions delivered, and the approach used to measure data leakage and student knowledge before and after the lessons. The Section 4 presents the key findings from the surveys and data leakage analysis, highlighting improvements in smartphone privacy knowledge and behavioural changes post-intervention. In Section 5, we analyse these findings in the context of prior research, exploring the implications for educational practices and policies. Finally, Section 6 summarizes the key takeaways and suggests directions for future research, emphasizing the importance of targeted interventions in enhancing cyber safety for adolescents.

2. Related Work

2.1. Knowledge of Smartphone Privacy Settings

Male and female attitudes towards privacy often diverge due to various factors, including early exposure to technology, societal expectations, and career aspirations. Males are frequently encouraged to engage with programming, hardware, and gaming from a young age, which contributes to their advanced technological skills and confidence in these areas [13]. In contrast, females typically develop proficiency in communication tools, social media, and creative applications, favouring technology that enhances productivity and collaboration [14]. These differences influence how each gender approaches online privacy, data sharing, and the protection of personal information.

Males' higher likelihood of pursuing STEM (science, technology, engineering, mathematics) careers further cultivates their technological expertise [15]. On the other hand, females often gravitate towards fields such as education, healthcare, and social sciences, where the use of technology is more focused on application rather than development [16,17].

This divergence in career interests may impact their knowledge of and attitudes toward technical privacy settings.

Males also tend to report higher confidence in performing technological tasks [18,19], which could contribute to a perception of greater technological proficiency among younger males. This confidence, potentially bolstered by frequent engagement with video games and other technology-intensive activities, might lead to higher levels of perceived knowledge [20]. However, despite having comparable abilities, females may exhibit lower self-efficacy in traditionally male-dominated areas [21,22]. This discrepancy in confidence can result in females perceiving themselves as less knowledgeable, even if their actual skills are equivalent.

The perception that technology settings are a "male activity" could further influence females' enthusiasm for learning about technology [19], particularly if smartphone privacy settings are seen as more technical [23]. This societal bias might discourage females from engaging deeply with privacy settings, reinforcing the gender gap in perceived technological expertise. Based on these gender differences that are likely to result in differences for adolescent male and female knowledge of smartphone settings, we propose the following hypothesis:

Hypothesis 1: *Adolescent males are expected to have higher levels of knowledge of smartphone privacy settings than adolescent females.*

2.2. Attitude towards Smartphone Privacy Settings

Male and female teenagers exhibit differing levels of privacy practices, influenced by factors such as social acceptance and peer approval. Male teenagers generally perceive lower risks related to online privacy, often prioritizing convenience over security. This behaviour is reflected in their tendency to share personal information more freely on social media, potentially leading to less effective use of privacy settings [24]. Social acceptance and peer approval often take precedence over privacy concerns for males, which may cause them to underestimate the potential consequences, such as privacy breaches or data misuse [10].

In contrast, female teenagers tend to be more proactive in protecting their privacy. They are more likely to utilize simple privacy protection settings on social media (e.g., untagging photos), limit the sharing of personal information, and carefully select their online interactions [24,25]. Females are more inclined to manage their online social media presence with greater caution, sharing content primarily with trusted friends and family while being vigilant about what they disclose [24,26]. This contrasts with males, who are more inclined to share personal data with third-party apps, often prioritizing convenience and enhanced functionality over informed consent [27]. The gender differences for attitudes towards privacy may translate into differences in attitudes for male and female adolescent attitudes towards smartphone settings. We propose the following hypothesis:

Hypothesis 2: *Adolescent females are expected to have a more cautious attitude towards smartphone privacy settings than adolescent males.*

2.3. Concerns About Smartphone Privacy

While female teenagers are more likely to seek privacy-enhancing tools and value informed consent [28], their strong motivation for social connection can sometimes lead to compromises in privacy for the sake of maintaining online interactions [29]. Despite their general caution, the desire for social connection may cause them to occasionally overlook privacy risks. Females are more likely than males to adopt protective practices, such as limiting contact with strangers and using stronger privacy settings, due to a heightened awareness of issues like online harassment [30]. Additionally, females often prioritize physical safety in their online behaviour, employing strategies like using pseudonyms

and avoiding the sharing of real-time locations to mitigate potential risks [31]. Given the anticipated gender differences, where adolescent females are more likely to worry about smartphone privacy due to concerns related to physical safety and location risks than adolescent males, we propose the following hypothesis:

Hypothesis 3: *Adolescent females are expected to have higher levels of concerns about smartphone privacy than adolescent males.*

2.4. Privacy Paradox

The privacy paradox describes a phenomenon where individuals express concerns about their privacy yet engage in behaviours that contradict these concerns, often compromising their privacy in online environments [32]. This paradox is especially evident in digital spaces, where users claim to value their privacy but still share personal information freely, use weak passwords, or neglect to adjust privacy settings [33].

For adolescents, the privacy paradox is particularly pronounced due to their still-developing capacity to fully comprehend long-term consequences and risks [34]. Adolescents are more susceptible to impulsive behaviour, especially in the context of peer interactions or social validation [9]. This impulsivity can lead to sharing personal information online without fully considering the associated privacy risks [9]. While they may be aware of privacy concerns such as cyberbullying, identity theft, or unwanted attention and express anxiety about these threats, they frequently engage in risky behaviours like sharing personal details, using weak passwords, or failing to log out of shared devices [35].

Gender differences are likely to influence how the privacy paradox manifests among adolescents. These differences are shaped by variations in socialization, risk perception, and the ways in which boys and girls interact with technology [24]. Males, who may be less concerned about privacy breaches related to physical safety, such as stalking or harassment, often focus more on cybersecurity threats like hacking and identity theft, employing tools like two-factor authentication to protect themselves [36].

Females, on the other hand, typically perceive higher risks regarding privacy, especially concerning personal data being accessed and misused by third parties, potentially leading to safety issues and harassment [24,37]. However, despite their heightened awareness of these risks, they may paradoxically expose themselves to harm by freely using real names and profiles across platforms for social connection purposes [38]. This behaviour exemplifies the privacy paradox, where high-risk perception does not always translate into protective actions. Given these observations, where adolescent males are likely to have higher levels of knowledge of smartphone settings and a less cautious attitude to smartphone privacy settings than adolescent females, we propose the following hypothesis:

Hypothesis 4: *The privacy paradox between levels of knowledge of smartphone privacy settings and levels of attitudes toward smartphone privacy settings will be more pronounced for adolescent males than for adolescent females.*

2.5. Targeted Educational Intervention

Addressing gender differences in privacy through inclusive education for technological skills may help bridge gaps in knowledge, attitudes, and concerns related to smartphone privacy settings by encouraging all teenagers to explore and develop a diverse range of competencies. While many studies have applied theories such as communication privacy management [39] and the privacy calculus model [40,41], this study leverages PMT to examine how threat and coping appraisals influence protective behaviour. PMT suggests that individuals are motivated to protect themselves when they perceive a threat as both severe and credible (perceived threat) and believe they have the capability and efficacy to respond effectively (perceived efficacy) [12]. This framework is particularly relevant

to smartphone privacy, as it helps identify the factors that may deter individuals from engaging in protective behaviours.

Linking PMT with educational programs has proven critical for enhancing cyber privacy protection [42]. Research indicates that educational interventions can significantly improve children's digital skills and self-efficacy [43]. However, much of the current cybersecurity privacy messaging lacks proper evaluation regarding its effectiveness in being learned, applied, and the actual safety it provides [44] and differences between genders. Understanding how PMT can address gender disparities in privacy attitudes is essential, as these differences highlight the need for tailored education that addresses specific concerns and behaviours. Considering that targeted educational interventions are expected to reduce gender differences in privacy threat perceptions and efficacy based on PMT, we propose the following hypothesis:

Hypothesis 5: *A targeted educational intervention is expected to reduce gender disparities in knowledge, attitudes, and concerns regarding smartphone privacy among adolescent males and females.*

3. Materials and Methods

3.1. Procedure

The presenters delivered a one-off hands-on lesson of approximately one hour duration in five Australian high schools. The lesson demonstrated how smartphones can inadvertently share sensitive information, how easily this information can be captured by others, and how students can manage their devices to prevent such data leaks. The lesson included practical demonstrations where students learned how to turn off Bluetooth, switch off Wi-Fi, change their Bluetooth name, and disable location services.

Before and after the lesson, students were invited to complete a short online survey to assess changes in their awareness of online safety. Participation was voluntary and not targeted by gender or ethnicity. Each student received a QR code linked to a unique identifier for survey access. All participants received an AUD 20 gift card, regardless of survey completion. Ethical approval was obtained from the university's Human Research Ethics Committee (Protocol No. H23489). Of 574 responses, 79 were excluded due to incomplete surveys, leaving 495 valid responses for analysis.

3.2. Participants

To confirm the validity of the survey responses, the data were analysed to identify duplicate participant IDs among the high school students who had shared the same QR code. To differentiate between unique responses, different IP addresses suggested that the survey was accessed from different devices. Gender, age, and high school information were also used as unique identifiers to differentiate between responses.

Where all information was valid between the responses and there was a time difference that aligned to before and after the presentation, these were classified as "Before" responses, where the timeframe aligned to prior to the cyber safety lesson, and "after" responses, where the timeframe aligned to after the cyber safety lesson. The dataset included 376 "before" responses and 100 "after" responses.

Table 1 presents the sociodemographic statistics for the before and after groups, including age, gender, school, and device ownership. Gender distribution remained relatively consistent, with males comprising 41.2% of the before group and 47.0% of the after group, and females making up 58.8% and 53.0%, respectively. Participants ranged in age from 13 to 19 years, with a mean age of 14.96 years (SD = 1.07) in the before group and 14.76 years (SD = 0.94) in the after group. Most students were between 14 and 16 years old (>90% in both groups), and 93% owned a smartphone. Smartphones accounted for 44.5% of all devices owned, with 42% of students indicating they owned only one type of digital device.

Table 1. Summary of sociodemographic statistics for the before and after groups.

		Before		After	
		n	**%**	***n***	**%**
Age	13	8	2.1	1	1.0
	14	155	41.2	54	54.0
	15	88	23.4	15	15.0
	16	100	26.6	28	28.0
	17	19	5.1	2	2.0
	18	5	1.3	-	-
	19	1	0.3	-	-
Gender	Male	155	41.2	47	47.0
	Female	221	58.8	53	53.0
Devices Owned	Smartphone	342	44.5	92	43.2
	Smartwatch	113	14.7	34	16.0
	iPad	125	16.3	34	16.0
	Android Tablet	25	3.3	5	2.3
	Fitness Tracker	52	6.8	14	6.6
	Clothing with "Smart Tags"	7	0.9	5	2.3
	Dumb Phone	15	2.0	2	0.9
	None	3	0.4	27	12.7
	Other	87	11.3	92	43.2
Number of Devices Owned	0	3	0.8	2	1.9
	1	159	42.3	39	37.5
	2	99	26.3	26	25.0
	3	70	18.6	24	23.1
	4	35	9.3	12	11.5
	5	7	1.9	1	1.0
	6	1	0.3	2	1.9
	7	2	0.5	39	37.5

3.3. Measures

Knowledge of smartphone privacy settings was measured using 13 questionnaire items rated on a four-point Likert scale from one (Strongly disagree) to four (strongly agree). The questions asked students about their knowledge of smartphone names, location services, Bluetooth, Wi-Fi, IP address, and SSID settings. This questionnaire received a total of 476 responses from students. There were 375 "before" responses and 101 "after" responses.

Attitudes towards smartphone privacy settings were measured using 6 questionnaire items rated on a four-point Likert scale from one (strongly disagree) to four (strongly agree). The questions asked students about their attitude towards accessing public Wi-Fi's, using their real name, switching off location services, Bluetooth and Wi-Fi, and whether or not they worry about their cybersecurity. This questionnaire received a total of 476 responses from students. There were 367 "before" responses and 96 "after" responses.

Concerns about smartphone privacy was measured using 13 questionnaire items rated on a four-point Likert scale from one (strongly disagree) to four (strongly agree). The questions were grouped into four main privacy concern categories with an additional two questions for each category outlining whether their concern was heightened "if the person monitoring their privacy was a stranger" and "if the monitoring was happening without their awareness". A total of 448 responses were included in the analysis. There were 358 "before" responses and 90 "after" responses.

3.4. Data Analysis

Data analysis was conducted using IBM SPSS Statistics for Windows, Version 26. To explore intergroup differences in responses before and after the intervention, as well as potential gender-related variances, we employed *t*-tests and one-way ANOVA. Data were collected via an online survey tool, extracted, and tabulated using Microsoft Excel Version 2406 (Microsoft Corporation, Redmond, WA, USA). An independent sample *t*-test was used to compare gender differences and evaluate before and after intervention outcomes. The mean (M) and standard deviation (SD) values were also calculated. Additionally,

analysis of variance (ANOVA) was used to assess the magnitude of changes within each group, determining t-values (t) and statistically significant *p*-values (*p*). Eta-squared (η^2) was calculated as a measure of effect size. Tukey's post hoc test was conducted to facilitate pairwise comparisons between mean scores, particularly in assessing differences between before and after intervention results across schools. Statistical significance was set at * $p < 0.05$ and ** $p < 0.01$.

4. Results

Figure 1 presents the mean differences between males and females for each question set before and after the intervention. Females had the highest overall mean scores for the concerns about smartphone privacy questions, while males had the highest mean scores for knowledge of smartphone privacy settings and attitude towards smartphone privacy settings.

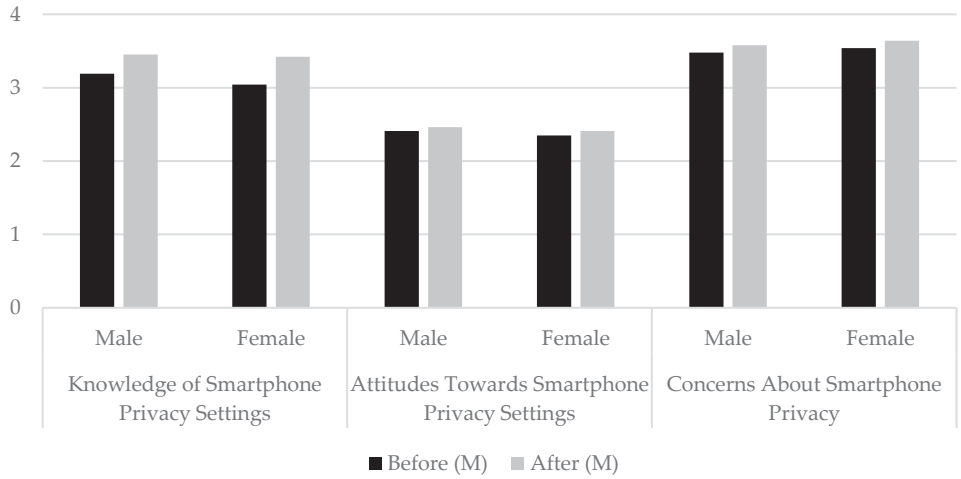

Figure 1. Gender before and after mean results for each survey questionnaire.

Table 2 presents the overall mean scores for each question set, comparing gender results for the before and after groups. Notably, all mean scores for both males and females increased in the after group compared to the before group.

Table 2. Overall mean statistics for each question set relating to each gender for the before and after groups.

		Before			After		
		n	**M**	**SD**	***n***	**M**	**SD**
Knowledge of Smartphone Privacy Settings	Male	147	3.19	0.53	46	3.45	0.44
	Female	213	3.04	0.55	51	3.42	0.45
Attitudes towards Smartphone Privacy Settings	Male	142	2.41	0.65	43	2.46	0.65
	Female	210	2.35	0.71	50	2.41	0.87
Concerns About Smartphone Privacy	Male	138	3.48	0.54	38	3.58	0.51
	Female	206	3.54	0.56	49	3.64	0.46

4.1. Knowledge of Smartphone Privacy Settings

Table 3 shows the mean and standard deviation for the knowledge of smartphone privacy settings questions based on both the male and female responses for the before and after groups. For the before group, males scored higher than females in 11 out of 13 knowledge of smartphone privacy settings questions. For the after group, males scored higher on seven out of 13 knowledge of smartphone privacy settings questions, which indicated a reduction in the differences between male and female knowledge levels post

the cyber safety lesson. The results indicate that both males and females possess a generally high level of knowledge regarding smartphone settings, with mean scores predominantly above 3. This suggests that both genders had a solid understanding of essential smartphone functionalities post the cyber safety intervention lesson.

Table 3. Mean results for knowledge of smartphone settings for male and female before and after groups.

			Before			After		
	Question	**Gender**	***n***	**M**	**SD**	***n***	**M**	**SD**
1	I know how to change my name on my smartphone	Male Female	147 213	3.39 3.41	0.73 0.76	46 51	3.76 3.76	0.43 0.47
2	I know how to switch off my location services	Male Female	147 213	3.48 3.50	0.66 0.69	46 51	3.63 3.67	0.57 0.59
3	I know what an SSID is	Male Female	147 213	2.53 2.01	0.97 0.93	46 51	2.93 2.47	0.93 0.99
4	I know how to switch off my Wi-Fi connection	Male Female	147 213	3.79 3.74	0.46 0.54	46 51	3.76 3.73	0.43 0.57
5	I know what my Wi-Fi is broadcasting	Male Female	147 213	3.07 2.77	0.87 0.93	46 51	3.50 3.35	0.66 0.72
6	I know what Bluetooth advertisement is	Male Female	147 213	2.90 2.81	0.89 0.92	46 51	3.30 3.29	0.73 0.78
7	I know how to switch off my Bluetooth connection	Male Female	147 213	3.76 3.71	0.49 0.55	46 51	3.67 3.75	0.47 0.48
8	I know how to prevent apps from tracking my activities	Male Female	147 213	3.31 3.05	0.70 0.84	46 51	3.43 3.49	0.69 0.67
9	I know how to prevent IP Address Tracking	Male Female	147 213	2.67 2.50	0.87 0.99	46 51	3.00 3.12	0.84 0.89
10	I know what my Bluetooth device is Broadcasting	Male Female	147 213	2.91 2.74	0.87 0.97	46 51	3.33 3.27	0.70 0.87
11	I know how to change my Bluetooth name	Male Female	147 213	3.21 3.08	0.85 0.89	46 51	3.52 3.57	0.66 0.67
12	I know how to manage my smartphone settings	Male Female	147 213	3.48 3.34	0.58 0.70	46 51	3.59 3.57	0.54 0.54
13	I know that my smartphone continually leaks information	Male Female	147 213	3.02 2.88	0.88 0.87	46 51	3.48 3.43	0.59 0.61

Before intervention, both male and female participants displayed a reasonably high level of knowledge about smartphone settings, with males generally scoring slightly higher across most questions. Males showed a higher familiarity with concepts like SSID (M = 2.53, SD = 0.97) and what their Wi-Fi is broadcasting (M = 3.07, SD = 0.87) compared to females (SSID: M = 2.01, SD = 0.93; Wi-Fi broadcasting: M = 2.77, SD = 0.93). Both genders had the highest confidence in knowing how to switch off their Wi-Fi connection (males: M = 3.79, SD = 0.46; females: M = 3.74, SD = 0.54). On average, males reported higher knowledge than females in nearly all aspects of smartphone settings. The most significant gaps were observed in technical areas, such as knowing what an SSID is and understanding what their Wi-Fi or Bluetooth devices broadcast.

After intervention, knowledge levels improved across all items for both genders, with noticeable increases in understanding complex concepts like preventing IP address tracking (males: M = 3.00, SD = 0.84; females: M = 3.12, SD = 0.89) and what Bluetooth devices broadcast (males: M = 3.33, SD = 0.70; females: M = 3.27, SD = 0.87). The gap between male and female knowledge narrowed after the intervention, particularly in areas like

changing Bluetooth names (males: M = 3.52, SD = 0.66; females: M = 3.57, SD = 0.67) and managing smartphone settings (males: M = 3.59, SD = 0.54; females: M = 3.57, SD = 0.54). Both genders saw increases in their mean scores, with the most substantial gains in areas where knowledge was initially lower, such as understanding SSID and IP address tracking.

The variability in responses was generally higher among females than males before the intervention, particularly in knowledge areas that were more technical, such as SSID and what their devices broadcast via Bluetooth or Wi-Fi. This suggests a broader range of understanding within the female group. Post the intervention, standard deviations generally decreased, indicating a more consistent level of knowledge across participants, with some areas showing significant convergence between males and females. For example, knowledge about preventing IP address tracking saw a notable reduction in variability for both genders, suggesting the intervention was effective in standardizing this understanding.

A two-sample *t*-test and ANOVA means analysis were performed to compare knowledge of smartphone privacy settings questions results between males (n = 147) and females (n = 213) in the before group. There was a significant difference in the means for question three relating to students "knowing what an SSID is" between males (M = 2.53, SD = 0.97) and females (M = 2.01, SD = 0.93); t(304) = 5.038, p = 0.000, η^2 = 0.067. There was a significant difference in the means for question five relating to students "knowing what their Wi-Fi is broadcasting" between males (M = 3.07, SD = 0.87) and females (M = 2.77, SD = 0.93); t(358) = 3.083, p = 0.002, η^2 = 0.026. There was a significant difference in the means for question eight relating to students "knowing how to prevent apps from tracking their activities" between males (M = 3.31, SD = 0.70) and females (M = 3.05, SD = 0.84); t(358) = 3.094, p = 0.002, η^2 = 0.026. There was a significant difference in the means for question 12 relating to students 'knowing how to manage their smartphone settings' between males (M = 3.48, SD = 0.58) and females (M = 3.34, SD = 0.70); t(358) = 2.003, p = 0.046, η^2 = 0.011.

A two-sample *t*-test and ANOVA means analysis were performed to compare knowledge of smartphone privacy settings questions results between males (n = 46) and females (n = 51) in the after group. There was only one significant difference in the means post the cybersecurity lesson and that was for question three relating to "students knowing what an SSID is" between males (M = 2.93, SD = 0.93) and females (M = 2.47, SD = 0.99); t(95) = 2.386, p = 0.019, η^2 = 0.056. There was no significant difference in the means for the remaining knowledge of smartphone privacy settings questions post the cybersecurity lesson, which indicates the cybersecurity lesson was effective in increasing both male and female knowledge of smartphone privacy settings.

The results indicate that Hypothesis 1 was supported, as males were found to have a higher level of knowledge of smartphone privacy settings than females. The intervention appears to have effectively elevated knowledge in these more technical aspects, reducing the gender gap observed before the intervention.

4.2. Attitudes towards Smartphone Privacy Settings

Table 4 presents the mean and standard deviation for responses to the attitude towards smartphone privacy settings questions, comparing male and female responses before and after the cyber safety lesson. In both the before and after groups, males and females scored three out of six questions higher. Females reported higher scores than males in the following areas: regularly switching off location services, regularly switching off Wi-Fi, in the after group only, switching off Bluetooth radio regularly and, in the before group only, concerns about smartphone security not being adequately addressed by their device.

Table 4. Mean results for attitude towards smartphone settings questions for male and female before and after groups.

			Before			After		
	Question	**Gender**	***n***	**M**	**SD**	***n***	**M**	**SD**
1	I don't use public Wi-Fis whenever they are available	Male	142	2.77	1.03	43	2.88	1.00
		Female	210	2.60	0.95	50	2.62	1.10
2	I don't use my real name as my smartphone name	Male	142	2.35	0.98	43	2.79	0.97
		Female	210	2.10	0.98	50	2.40	0.97
3	I switch off my location services regularly	Male	142	2.32	0.89	43	2.21	0.89
		Female	210	2.38	0.89	50	2.28	0.97
4	I switch off my Wi-Fi connection regularly	Male	142	2.15	0.85	43	2.07	0.74
		Female	210	2.22	0.92	50	2.30	1.04
5	I switch off my Bluetooth radio regularly	Male	142	2.43	0.96	43	2.28	0.80
		Female	210	2.31	0.95	50	2.40	1.01
6	I worry about smartphone security because my smartphone does not take care of it	Male	142	2.42	0.89	43	2.53	0.96
		Female	210	2.47	0.94	50	2.44	1.01

Before intervention, both males and females exhibited moderate levels of protective behaviours regarding smartphone use. Males were slightly more likely than females to avoid using public Wi-Fi (M = 2.77, SD = 1.03) and to not use their real name as their smartphone name (M = 2.35, SD = 0.98). However, neither group consistently engaged in practices like switching off location services or Bluetooth regularly, with mean scores around 2.3 for both genders. In comparison, in the post-intervention group, there were modest improvements in privacy protective behaviours. Both genders reported slight increases in not using their real name as their smartphone name (Males: M = 2.79, SD = 0.97; Females: M = 2.40, SD = 0.97). However, behaviours such as regularly switching off Wi-Fi and location services remained relatively unchanged, indicating persistent habits that may be harder to shift.

Males generally reported slightly higher protective behaviours than females across most questions before the intervention, particularly in avoiding public Wi-Fi and not using their real name as a smartphone name. The differences were small but consistent. Both genders showed small improvements in their protective behaviours, with males continuing to report slightly higher levels of these behaviours than females in the after group. The most notable improvement was in the practice of not using real names for smartphone identification, especially among males (M = 2.79, SD = 0.97).

The variability in responses was similar between males and females in the before group, with standard deviations indicating a moderate spread in protective behaviours. The highest variability was observed in the use of public Wi-Fi and the switching off of Bluetooth, suggesting diverse attitudes and practices within each gender group. Standard deviations remained relatively stable post-intervention, with no significant reductions in variability. This indicates that while mean behaviours improved slightly, the range of attitudes and practices among participants did not become more uniform.

A two-sample *t*-test and ANOVA means analysis were performed to compare knowledge of smartphone privacy settings questions results for males (n = 142) and females (n = 210) in the before group. There was a significant difference in the means for question two relating to whether students "didn't use their real name as their smartphone name" between males (M = 2.35, SD = 0.98) and females (M = 2.10, SD = 0.98); t(350) = 2.339, p = 0.020, η^2 = 0.048. There were no statistically significant differences for the attitudes towards smartphone privacy settings questions for after group between males and females indicating the educational intervention led to more consistent levels of privacy attitudes post the cybersecurity lesson.

The results indicate that Hypothesis 2 was not supported, as both males and females exhibited cautious attitudes toward smartphone settings, albeit in different areas. In the preintervention group, males showed more caution in technical aspects, while females

were more cautious regarding location privacy and general security concerns, aligning with expected gender differences in knowledge of smartphone settings and concerns about privacy.

The results supported Hypothesis 4, as the results underscore gender differences in knowledge and attitudes toward smartphone privacy and security, leading to increased privacy paradox for males. While both genders expressed moderate concern, males tended to score slightly higher on most measures. However, mean attitude scores were lower than knowledge scores, indicating a gap between awareness and behaviour. In the post-intervention group, this gap narrowed for means for both genders, with the difference between male and female scores decreasing from 0.09 to 0.02, though males continued to exhibit larger gaps between knowledge and attitudes.

4.3. Concerns About Smartphone Privacy

Table 5 shows the mean and standard deviation for concerns about smartphone privacy questions for males and females within the before and after group. Females scored higher means than males across the majority of the questions in the before and after group. Both male and female respondents exhibited a strong awareness of privacy concerns regarding their location data, with average scores consistently around 3.5 on a scale of one to four.

Table 5. Mean results for concerns about smartphone privacy for male and female before and after groups.

			Before			After		
	Question	**Gender**	***n***	**M**	**SD**	***n***	**M**	**SD**
1.	I would be concerned if someone knew where I was at any particular point in time	Male Female	138 206	3.24 3.23	0.80 0.80	38 49	3.47 3.49	0.65 0.65
1.1	I would be more concerned if that someone was a stranger	Male Female	138 206	3.49 3.64	0.70 0.62	38 49	3.55 3.69	0.69 0.51
1.2	I would be even more concerned if I didn't know it was happening	Male Female	138 206	3.54 3.59	0.62 0.65	38 49	3.61 3.67	0.64 0.47
2	I would be concerned if someone knew where I was at the same time every day	Male Female	138 206	3.43 3.44	0.66 0.70	38 49	3.55 3.59	0.60 0.54
2.1	I would be more concerned if that someone was a stranger	Male Female	138 206	3.56 3.66	0.64 0.64	38 49	3.61 3.71	0.50 0.50
2.2	I would be more concerned if I didn't know it was happening	Male Female	138 206	3.57 3.60	0.60 0.66	38 49	3.55 3.69	0.69 0.47
3	I would be concerned if someone knew what my regular travel route to school was	Male Female	138 206	3.25 3.34	0.74 0.80	38 49	3.55 3.55	0.65 0.58
3.1	I would be more concerned if that someone was a stranger	Male Female	138 206	3.50 3.60	0.66 0.69	38 49	3.68 3.71	0.47 0.50
3.2	I would be more concerned if I didn't know it was happening	Male Female	138 206	3.52 3.57	0.61 0.66	38 49	3.58 3.65	0.60 0.48
4	I would be concerned if someone was able to plot my location on a Google map at a particular point in time	Male Female	138 206	3.49 3.56	0.64 0.65	38 49	3.53 3.59	0.65 0.57
4.1	I would be concerned if someone could plot my location on a Google map over a period of time	Male Female	138 206	3.53 3.55	0.63 0.69	38 49	3.61 3.67	0.60 0.47
4.2	I would be more concerned if that someone was a stranger	Male Female	138 206	3.54 3.64	0.64 0.65	38 49	3.58 3.67	0.55 0.52
4.3	I would be more concerned if I didn't know it was happening	Male Female	138 206	3.56 3.61	0.62 0.66	38 49	3.63 3.61	0.54 0.61

The highest concerns were related to the involvement of strangers and the lack of knowledge about being tracked, with males and females both expressing significant unease about these scenarios (e.g., "I would be more concerned if that someone was a stranger": males M = 3.49, SD = 0.70; females M = 3.64, SD = 0.62). There was a slight increase in concern levels, post-intervention, especially among females. For example, female concern

about someone knowing where they were at any particular point in time increased slightly (M = 3.49, SD = 0.65), as did concern about strangers having access to their location data over time (M = 3.67, SD = 0.47). Males also showed slight increases, but the changes were generally smaller.

The means indicate that females were generally more concerned about location privacy than males, particularly when considering scenarios involving strangers or unknown tracking in the before group. For instance, females reported higher concern about a stranger knowing their location (M = 3.64) compared to males (M = 3.49). Both genders exhibited small increases in concern levels post-intervention, with females maintaining slightly higher concern levels than males across most variables. Notably, the concern about location tracking by strangers remained one of the highest-rated items for both groups (males: M = 3.58; females: M = 3.67).

Standard deviations across the board were relatively low in the before group, suggesting a consensus among participants regarding their concerns about location privacy. The highest variability was observed in responses to concerns about what a stranger might know, indicating some differences in perception among participants. The standard deviations remained stable post-intervention, with only minor changes, indicating that the intervention did not significantly alter the distribution of responses. Females, in particular, showed a slight decrease in variability, suggesting a more uniform concern level post-intervention.

A two-sample *t*-test and ANOVA means analysis were performed to compare concerns about smartphone privacy questions results for males (n = 138) and females (n = 206) in the before group. There was a significant difference in the means for the second question in the first category of whether students "would be more concerned if a stranger knew where they were at any particular point in time" between males (M = 3.49, SD = 0.70) and females (M = 3.64, SD = 0.62); $t(271) = -2.015$, $p = 0.045$, $\eta^2 = 0.048$. There was no significant difference in the means for any of the other questions in the male and female groups for concerns about smartphone privacy. There were no statistically significant differences for the concerns about smartphone privacy questions for after group between males and females, indicating the educational intervention had some effect on consistency in attitude towards privacy post the cybersecurity lesson.

The results support Hypothesis 3, with females exhibiting higher levels of concern about smartphone privacy settings, especially related to personal safety and strangers, compared to males. The intervention successfully increased perceived threats for both genders, narrowing the preintervention gender gap.

The results also supported Hypothesis 5, as the targeted educational intervention reduced gender differences in knowledge, attitudes, and concerns regarding smartphone privacy. Post-intervention, responses from male and female adolescents were more consistent across all question sets, suggesting that educational interventions can effectively reduce gender disparities in smartphone privacy risk.

5. Discussion

5.1. Knowledge of Smartphone Privacy Settings

The results revealed that, prior to the intervention, females reported lower knowledge of smartphone settings compared to males. However, after the cyber safety lessons, both genders exhibited significant improvements in knowledge, leading to more consistent levels across genders. This finding highlights the effectiveness of the intervention in enhancing students' understanding of smartphone security settings, aligning with PMT [12], which suggests that perceived threat and efficacy can drive behavioural change.

Despite overall improvements in knowledge across both genders, technical aspects like SSIDs and broadcasting settings remained challenging, particularly for females. Males exhibited slightly better technical understanding, supporting previous findings on gender differences in confidence and technical knowledge [13]. The results suggest that targeted interventions may be necessary to bridge this gap. Interestingly, even males—who initially

reported higher confidence in their knowledge—demonstrated significant improvement, suggesting the value of continual reinforcement of PMT concepts for addressing ongoing behaviour change [12].

While the results showed that the intervention helped equalize knowledge between genders, the results also indicate that post-intervention, females rated their knowledge on par with males, highlighting that when equipped with equal knowledge, both genders can perform similarly. This suggests that prior to the lesson, females may have lacked confidence rather than ability. This finding aligns with previous research, which suggests that a lack of confidence may lead females to perceive themselves as less knowledgeable, even when their actual skills are comparable to males [21]. Social biases often portray men as being more proficient in technology and this is likely to contribute to females self-reporting lower confidence in technological skills [19,20].

5.2. Attitudes towards Smartphone Privacy Settings

Despite increased awareness, the persistence of behaviours like keeping Wi-Fi and Bluetooth enabled underscores the privacy paradox [32], where knowledge does not always translate into protective actions. These results emphasize the need for educational programs to not only increase knowledge but also actively change behaviours regarding smartphone privacy management. The gender differences observed in attitudes toward privacy management suggest that education should target specific concerns and behaviours. Programs should focus not only on increasing knowledge of smartphone privacy settings but also on shaping attitudes toward actively managing these settings.

Future research could explore the psychological or contextual factors that contribute to the persistence of risky behaviours, such as the reluctance to disable Wi-Fi or Bluetooth. Additionally, evaluating the impact of different educational interventions, such as gamified learning or peer-led discussions, could provide valuable insights into improving behaviours. Additionally, evaluating whether these risky behaviours correlate with actual privacy breaches would offer a more comprehensive understanding of the practical effectiveness of these educational interventions in promoting cyber safety. Although targeted interventions resulted in modest improvements in behaviours, such as avoiding real-name use for smartphone identification, other behaviours—like regularly disabling location services or Wi-Fi connections—remained more resistant to significant change. This suggests that ongoing education is crucial to not only raise awareness but also to motivate more consistent behavioural changes. Tailoring educational messages to address specific misconceptions or barriers could significantly enhance the effectiveness of these interventions.

The results underscore gender differences in the privacy paradox [32]. While both genders expressed moderate concern, males tended to score slightly higher on most measures. However, mean attitude scores were lower than knowledge scores, indicating a gap between awareness and behaviour. This aligns with previous research showing that while females perceive greater privacy risks, especially related to personal data misuse and harassment, their attitudes may not always translate into protective actions [24,37], and males have a tendency to share personal information more freely on social media, potentially leading to less effective use of privacy settings [24].

5.3. Concerns About Smartphone Privacy

The findings highlight gender differences in privacy concerns related to location data. Both genders showed moderate concern about their whereabouts being known, with males slightly more concerned overall, but females expressing greater concern when strangers were involved, aligning with other research on gender and physical safety concerns [30]. This concern persisted across scenarios, such as being tracked daily or having their travel route known.

Both genders consistently demonstrated awareness of location tracking risks. Females were particularly worried about being tracked via Google Maps, especially by strangers or without their knowledge. These results suggest that females' concerns are often linked

to personal safety, emphasizing the need for education that addresses both privacy and physical security risks.

The findings highlight the need for tailored cyber education that not only addresses knowledge gaps but also focuses on motivating protective behaviours, particularly where privacy concern gaps persist. Future interventions should prioritize educating adolescents about location privacy, especially in relation to personal safety concerns for being tracked by unknown individuals. While self-reported concern levels were already high, targeted education can further increase awareness and promote safer behaviours. Future interventions should not only raise awareness but also teach practical strategies for managing location privacy, like effective use of privacy settings and understanding the implications of location sharing. Further research could explore whether increased concern leads to more cautious use of location services, the factors driving higher concern among females, and how peer influence shapes attitudes toward location privacy. This could help in designing more impactful educational programs. Teaching practical strategies for managing location privacy and emphasizing the implications of location sharing will be essential for fostering safer digital practices among adolescents.

6. Conclusions

This study contributes to the understanding of how gender differences influence privacy behaviours among adolescents and underscores the need for tailored educational interventions to address these disparities. Our study found that before the cyber safety intervention, females reported lower knowledge of smartphone settings than males. After the lesson, this gap narrowed, with both genders demonstrating more consistent understanding. Males still had a slight edge in technical areas like SSIDs and broadcasting settings, highlighting the need for targeted interventions to support female students in these areas. Both genders showed lower attitudes towards smartphone privacy compared to knowledge, with males displaying the largest gap, reflecting the privacy paradox. Females expressed greater concern regarding location privacy, especially when tracked by unknown individuals, indicating that while both genders are aware of risks, females perceive them more acutely.

The results suggest that targeted educational programs can effectively enhance adolescents' knowledge, attitudes, and concerns about smartphone privacy, particularly in technical areas where gender gaps exist. The narrowing of the knowledge gap across both genders post-intervention underscores the importance of such programs in building essential digital literacy skills, though persistent differences highlight the need for continued, focused efforts. Educational interventions should not only address the knowledge gap but also focus on changing behaviours related to privacy. Adolescents need practical skills to manage privacy settings, understand data-sharing risks, and recognize phishing attempts. By acknowledging gender differences, digital literacy programs can be tailored to meet distinct needs. These differences are influenced by education, cultural norms, access to technology, and personal interests.

These findings highlight the potential for targeted educational interventions to significantly improve smartphone privacy knowledge and reduce risky behaviours among adolescents. Schools can incorporate cyber safety lessons into their curricula, supported by policies that mandate regular education on managing smartphone privacy settings and mitigating data leakage. The demonstrated reduction in data leakage post-lesson suggests real behavioural change, with implications for shaping broader government and advocacy programs. Tailored interventions, particularly for female students who were found to be more at risk, can further enhance the effectiveness of these initiatives. Ultimately, these efforts could lead to a generation of digitally literate adolescents better equipped to navigate online risks, informing both school policies and national cyber safety strategies.

A limitation of this study was the small size of the post-lesson survey group and the lack of long-term impact assessment. Future research should ensure that all prelesson survey participants also complete the post-lesson survey and conduct longitudinal studies

to evaluate the lasting effectiveness of cyber safety education. Another limitation is understanding existing levels of technology use and psychological or contextual factors that contribute to the persistence of certain risky behaviours, such as the reluctance to switch off Wi-Fi or Bluetooth. Further research might also examine whether these behaviours correlate with actual incidents of privacy breaches, offering a more comprehensive understanding of the effectiveness of these practices. Future investigations could shed light on persistent knowledge gaps, especially concerning SSIDs and technical aspects, and understanding students' prior technology experiences could provide valuable insights.

Author Contributions: Conceptualization, Y.A.-S.; methodology, Y.A.-S.; software, Y.A.-S.; validation, Y.A.-S.; formal analysis, J.M.; investigation, Y.A.-S.; resources, Y.A.-S.; data curation, Y.A.-S. and J.M.; writing—original draft preparation, J.M.; writing—review and editing, Y.A.-S.; visualization, J.M.; supervision, Y.A.-S.; project administration, Y.A.-S.; funding acquisition, Y.A.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded through the Australian eSafety Commissioner's Online Safety Grants Program.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Charles Sturt University Human Research Ethics Committee (protocol code H23489 and date of approval 4/4/2023).

Informed Consent Statement: I confirm that "Informed consent was obtained from all subjects involved in the study".

Data Availability Statement: The datasets presented in this article are not readily available because the study involved children. Requests to access the datasets should be directed to yalsaggaf@csu.edu.au.

Acknowledgments: The authors wish to thank Alan Ibbett and acknowledge his contribution to this study. This study builds on his earlier work for his doctoral thesis project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Taylor, P. Percentage of Teenagers in the United States Who Have Access to a Smartphone at Home as of October 2023, by Gender. 28 February 2024. Available online: https://www.statista.com/statistics/256501/teen-cell-phone-and-smartphone-ownership-in-the-us-by-gender/ (accessed on 14 August 2024).
2. Boerman, S.; Kruikemeier, S.; Borgesius, F. Exploring Motivations for Online Privacy Protection Behavior: Insights from Panel Data. *Commun. Res.* **2021**, *48*, 953–977. [CrossRef]
3. Smit, E.G.; Van Noort, G.; Voorveld, H.A.M. Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe. *Comput. Hum. Behav.* **2014**, *32*, 15–22. [CrossRef]
4. Baruh, L.; Secinti, E.; Zeynep, C. Online Privacy Concerns and Privacy Management: A Meta-Analytical Review. *J. Commun.* **2017**, *67*, 26–53. [CrossRef]
5. Chai, S.; Das, S.; Rao, H. Factors Affecting Bloggers' Knowledge Sharing: An Investigation across Gender. *J. Manag. Inf. Syst.* **2011**, *28*, 309–341. [CrossRef]
6. Richter, A.; Adkins, V.; Selkie, E. Youth Perspectives on the Recommended Age of Mobile Phone Adoption: Survey Study. *JMIR Pediatr. Parent.* **2022**, *5*, e40704. [CrossRef]
7. Chang, V.; Golightly, L.; Xu, Q.A.; Boonmee, T.; Liu, B.S. Cybersecurity for children: An investigation into the application of social media. *Enterp. Inf. Syst.* **2023**, *17*, 2188122. [CrossRef]
8. Radesky, J.; Weeks, H.M.; Schaller, A.; Robb, M.; Mann, S.; Lenhart, A. *Constant Companion: A Week in the Life of a Young Person's Smartphone Use*; Common Sense: San Francisco, CA, USA, 2023.
9. Romer, D. Adolescent risk taking, impulsivity, and brain development: Implications for prevention. *Dev. Psychobiol.* **2010**, *52*, 263–276. [CrossRef]
10. Youn, S.; Hall, K. Gender and Online Privacy Among Teens: Risk Perception, Privacy Concerns, and Protection Behaviors. *Cyberpsychol. Behav.* **2008**, *11*, 763–765. [CrossRef]
11. Dempsey, J.; Sim, G.; Cassidy, B. Designing for GDPR—Investigating Children's Understanding of Privacy: A Survey Approach, In Proceedings of the BCS-HCI'18: 32nd Human Computer Interaction Conference, Belfast, UK, 2–6 July 2018.
12. Rogers, R.W. A protection motivation theory of fear appeals and attitude change. *J. Psychol. Interdiscip. Appl.* **1975**, *91*, 93–114. [CrossRef]
13. Saunders, M.A. The Role of Video Game Play, Gender Roles, and Career Decision Self-Efficacy in Development of STEM Career Interests & Motivation. Doctoral Dissertations, Louisiana Tech University, Ruston, LA, USA, 2021.

14. Campbell, K. Gender and Technology: Social Context and Intersectionality. In *Handbook of Research in Educational Communications and Technology*; Bishop, M.J., Boling, E., Elen, J., Svihla, V., Eds.; Springer: Cham, Switzerland, 2020. [CrossRef]
15. Wang, M.; Degol, J.L. Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educ. Psychol. Rev.* **2017**, *29*, 119–140. [CrossRef]
16. Stewart-Williams, S.; Halsey, L.G. Men, women and STEM: Why the differences and what should be done? *Eur. J. Personal.* **2021**, *35*, 3–39. [CrossRef]
17. Carranza, E.; Das, S.; Kotikula, A. *GenderBased Employment Segregation: Understanding Causes and Policy Interventions*; World Bank: Washington, DC, USA, 2023.
18. Christensen, M.A. Tracing the Gender Confidence Gap in Computing: A Cross-National Meta-Analysis of Gender Differences in Self-Assessed Technological Ability. *Soc. Sci. Res.* **2023**, *111*, 102853. [CrossRef] [PubMed]
19. He, J.; Freeman, L. Are Men More Technology-Oriented Than Women? The Role of Gender on the Development of General Computer Self-Efficacy of College Students. *J. Inf. Syst. Educ.* **2019**, *21*, 672.
20. Marja, L.; Overå, S. Are There Differences in Video Gaming and Use of Social Media among Boys and Girls?-A Mixed Methods Approach. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6085. [CrossRef]
21. Robinson, K.A.; Perez, T.; White-Levatich, A.; Linnenbrink-Garcia, L. Gender Differences and Roles of Two Science Self-Efficacy Beliefs in Predicting Post-College Outcomes. *J. Exp. Educ.* **2022**, *90*, 344–363. [CrossRef]
22. Denejkina, A. Generative AI—Gender Gap Identified in Skills and Confidence. 25 August 2023. Available online: https://youthinsight.com.au/education/generative-ai-gender-gap-identified-in-skills-and-confidence/#:~:text=Here,%20men%20were%20more%20likely,62%20per%20cent%20of%20girls (accessed on 14 August 2024).
23. Sebastián-Tirado, A.; Félix-Esbrí, S.; Forn, C.; Sanchis-Segura, C. Are gender-science stereotypes barriers for women in science, technology, engineering, and mathematics? Exploring when, how, and to whom in an experimentally-controlled setting. *Front. Psychol.* **2023**, *14*, 1219012. [CrossRef]
24. Tifferet, S. Gender differences in privacy tendencies on social network sites: A meta-analysis. *Comput. Hum. Behav.* **2018**, *93*, 1–12. [CrossRef]
25. Gruzd, A.; Hernández-García, Á. Privacy Concerns and Self-Disclosure in Private and Public Uses of Social Media. *Cyberpsychol. Behav. Soc. Netw.* **2018**, *21*, 418–428. [CrossRef]
26. eSatefy Commissioner. State of Play—Youth, Kids and Digital Dangers, Australian Government. 3 May 2018. Available online: https://www.esafety.gov.au/sites/default/files/2019-10/State%20of%20Play%20-%20Youth%20kids%20and%20digital%20dangers.pdf (accessed on 14 August 2024).
27. Office of the Australian Information Commissioner. Australian Community Attitudes to Privacy Survey 2020. Australian Government. September 2020. Available online: https://www.oaic.gov.au/engage-with-us/research-and-training-resources/research/australian-community-attitudes-to-privacy-survey/australian-community-attitudes-to-privacy-survey-2020 (accessed on 14 August 2024).
28. Office of the Australian Information Commissioner. Australian Community Attitudes to Privacy Survey 2023, Australian Government. 8 August 2023. Available online: https://www.oaic.gov.au/engage-with-us/research-and-training-resources/research/australian-community-attitudes-to-privacy-survey/australian-community-attitudes-to-privacy-survey-2023 (accessed on 14 August 2024).
29. Savoia, E.; Harriman, N.W.; Su, M.; Cote, T.; Shortland, N. Adolescents' Exposure to Online Risks: Gender Disparities and Vulnerabilities Related to Online Behaviors. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5786. [CrossRef]
30. Livingstone, S.; Stoilova, M.; Nandagiri, R. *Children's Data and Privacy Online: Growing Up in a Digital Age. An Evidence Review*; London School of Economics and Political Science: London, UK, 2019.
31. Coopamootoo, K.; Ng, M. "Un-Equal Online Safety?" A Gender Analysis of Security and Privacy Protection Advice and Behaviour Patterns. In Proceedings of the 32nd USENIX Security Symposium, Anaheim, CA, USA, 9–11 August 2023.
32. Solove, D. The Myth of the Privacy Paradox. *Georg. Wash. Law Rev.* **2021**, *89*, 1. [CrossRef]
33. Svirsky, D. Why Do People Avoid Information About Privacy? *J. Law Innov.* **2021**, *2*, 2.
34. Hargittai, E.; Marwick, A. "What Can I Really Do?": Explaining the Privacy Paradox with Online Apathy. *Int. J. Commun.* **2016**, *10*, 21.
35. Quayyum, F.; Cruzes, D.S.; Jaccheri, L. Cybersecurity awareness for children: A systematic literature review. *Int. J. Child-Comput. Interact.* **2021**, *30*, 100343. [CrossRef]
36. Pratama, A.R.; Firmansyah, F.M. Until you have something to lose! Loss aversion and two-factor authentication adoption. *Appl. Comput. Inform.* 2021; *ahead-of-print*.
37. Dhir, A.; Torsheim, T.; Pallesen, S.; Andreassen, C.S. Do Online Privacy Concerns Predict Selfie Behavior among Adolescents, Young Adults and Adults? *Front. Psychol.* **2017**, *8*, 815. [CrossRef]
38. Kaarakainen, M.; Hutri, H. Participating with a Real Name, a Nickname or by Being Anonymous?—Anonymous and Identifiable Users' Skills and Internet Usage Habits, 2016. Available online: http://urn.fi/URN:ISBN:978-952-03-0307-5 (accessed on 1 August 2024).
39. Petronio, S.; Caughlin, J.P. Communication Privacy Management Theory: Understanding Families. In *Engaging Theories in Family Communication: Multiple Perspectives*; Braithwaite, D.O., Baxter, L.A., Eds.; Routledge: New York, NY, USA, 2006; pp. 35–49. [CrossRef]

40. Meier, Y.; Krämer, N.C. The Privacy Calculus Revisited: An Empirical Investigation of Online Privacy Decisions on Between- and Within-Person Levels. *Commun. Res.* **2024**, *51*, 178–202. [CrossRef]
41. Peng, Z. A privacy calculus model perspective that explains why parents sharent. *Inf. Commun. Soc.* **2023**, 1–24. [CrossRef]
42. Khan, N.F.; Ikram, N.; Murtaza, H.; Javed, M. Evaluating protection motivation based cybersecurity awareness training on Kirkpatrick's Model. *Comput. Secur.* **2023**, *125*, 103049. [CrossRef]
43. Lee, A.Y.; Hancock, J.T. Developing digital resilience: An educational intervention improves elementary students' response to digital challenges. *Comput. Educ. Open* **2023**, *5*, 100144. [CrossRef]
44. Finkelhor, D.; Jones, L.; Mitchell, K. Teaching privacy: A flawed strategy for children's online safety. *Child Abus. Negl.* **2021**, *117*, 105064. [CrossRef]

information

MDPI

Article

Efficient Cryptographic Solutions for Unbalanced Private Set Intersection in Mobile Communication

Qian Feng [1,2], Shenglong Du [1,2,*], Wuzheng Tan [1,2] and Jian Weng [1,2]

[1] College of Cyber Security, Jinan University, Guangzhou 510632, China; jnu.fengq@gmail.com (Q.F.); tanwuzheng@jnu.edu.cn (W.T.); cryptjweng@gmail.com (J.W.)
[2] Guangdong Key Laboratory of Data Security and Privacy Preserving, Guangzhou 510632, China
* Correspondence: jndsl@stu2021.jnu.edu.cn

Abstract: Private Set Intersection (PSI) is a cryptographic method in secure multi-party computation that allows entities to identify common elements in their datasets without revealing their private data. Traditional approaches assume similar-sized datasets and equal computational power, overlooking practical imbalances. In real-world applications, dataset sizes and computational capacities often vary, particularly in the Internet of Things and mobile scenarios where device limitations restrict computational types. Traditional PSI protocols are inefficient here, as computational and communication complexities correlate with the size of larger datasets. Thus, adapting PSI protocols to these imbalances is crucial. This paper explores unbalanced PSI scenarios where one party (the receiver) has a relatively small dataset and limited computational power, while the other party (the sender) has a large amount of data and strong computational capabilities. It introduces three innovative solutions for unbalanced PSI: an unbalanced PSI protocol based on the Cuckoo filter, an unbalanced PSI protocol based on single-cloud assistance, and an unbalanced PSI protocol based on dual-cloud assistance, with each subsequent solution addressing the shortcomings of the previous one. Depending on performance and security needs, different protocols can be employed for applications such as private contact discovery.

Keywords: multi-party computation; private set intersection; Cuckoo filter; cloud assistance; private contact discovery

1. Introduction

1.1. Background

In today's digital age, data privacy and security have become critically important issues worldwide. With technological advancements and explosive growth in data volumes, individuals and institutions face unprecedented challenges in protecting their privacy. Privacy computing technologies have emerged in response to these challenges, enabling the secure computation and analysis of data without exposing the details of personal information. This is crucial for driving data-driven innovation and services while safeguarding personal privacy and data protection.

The PSI is a key technique in the field of privacy computing. It allows two or more parties to identify the common elements in their datasets without revealing their private data. This technology is highly useful in multiple application scenarios, such as cross-institutional data cooperation, fraud detection, and private contact discovery, without compromising user privacy. It has been applied in various fields, including the genetic testing of fully sequenced human genomes [1], private contact discovery [2], and botnet detection [3].

Despite the robust privacy protection features of PSI, there is a significant limitation: the low efficiency in scenarios where dataset sizes are imbalanced. Traditional PSI schemes typically assume that participants have similarly sized datasets and computational capabilities, which is not always the case. Particularly, in scenarios where one party is a server with

a large amount of data and the other one is a mobile device with limited computational power, this imbalance is pronounced. Unbalanced PSI [2,4–6] was proposed to address this issue by optimizing algorithms and protocol designs to enhance efficiency and feasibility between participants with different computational powers and dataset sizes.

1.2. Motivation

To overcome the drawbacks of traditional PSI in scenarios with significant disparities in dataset sizes of the participants, this paper proposes a more suitable unbalanced PSI protocol based on Cuckoo filter. However, in the unbalanced PSI protocol based on Cuckoo filter, the client-side involves complex cryptographic operations and requires managing and processing a substantial amount of filter data (it will be demonstrated in the following sections). This poses a significant computational and storage burden on resource-constrained client devices, such as smartphones or other portable devices, limiting the practicality of these technologies and potentially affecting device performance and user experience.

In light of this, the second approach in this paper considers introducing cloud computing as a solution, to leverage its robust computational and storage capabilities to alleviate the load on client devices. By outsourcing part of the computation tasks to cloud servers, the workload of the client can be significantly reduced, thereby speeding up the entire set intersection process. However, outsourcing data processing tasks to the cloud environment introduces a new problem: the risk of collusion attacks. When the cloud server colludes with one of the participants, it could threaten the privacy security of the entire scheme.

To address this challenge, the paper proposes a third, more secure design aimed at thwarting potential collusion attacks while maintaining the efficiency and practicality of unbalanced PSI. This design incorporates multiple security technologies and strategies, effectively reducing the computational and storage pressure on client devices while ensuring data privacy in the cloud environment, even in the face of collusion threats. This is crucial for advancing the development and application of unbalanced PSI technologies, providing a secure, efficient, and practical framework for future privacy-preserving computations.

Moreover, the unbalanced PSI protocols proposed in this paper are particularly suited for private contact discovery applications. This scenario requires identifying and verifying common contacts between individuals or organizations under the premise of maintaining individual privacy, which is extremely important for social networking services, emergency response coordination, and business cooperation. Using the protocols proposed in this paper, users can securely and quickly identify common contacts without disclosing their complete contact lists. This not only enhances user privacy but also facilitates complex social network analyses and emergency contact networks, while avoiding security risks associated with data breaches or improper handling.

1.3. Main Work

Overall, this paper's research primarily addresses the shortcomings of traditional PSI protocols in unbalanced scenarios by proposing three unbalanced PSI protocols for different contexts. There is a progressive relationship between each protocol, with each new proposal improving upon the deficiencies of the previous one. Specifically, this paper introduces an unbalanced PSI protocol based on Cuckoo filter that resolves the traditional PSI protocol's issues with unbalanced scenarios, a single-cloud assisted unbalanced PSI protocol that transfers most computational and storage tasks from the client to the cloud, and a double-cloud assisted unbalanced PSI protocol that can resist collusion attacks. In practical applications, different schemes can be selected based on varying performance and security needs.

2. Related Works

2.1. Design Framework of PSI Protocol

2.1.1. Design Framework Based on Public Key Encryption

The basic idea behind early PSI is to encrypt data elements and then perform comparison operations on the encrypted data. The most widely used technique in this method is homomorphic encryption: the sender encrypts their dataset and sends it to the receiver. The receiver processes these ciphertexts using the properties of homomorphic encryption and returns the results to the sender. The sender then decrypts these results using their own private key to obtain the intersection of the datasets. This public-key-based method generally relies on three main security assumptions [7]:

1. Based on Diffie–Hellman (DH) theory: Meadows [8] used the DH key exchange mechanism, which is based on the discrete logarithm problem, to implement a PSI protocol. In contrast, Huberman [9] and his team explored the use of elliptic curve cryptography in PSI, noting its significant advantages in security and efficiency compared to traditional discrete logarithm-based PSI methods.
2. Based on the RSA assumption: DeCristofaro and others [10] developed a semi-honest PSI protocol using RSA blind signature technology based on the integer factorization problem. Another study [11] showed that PSI schemes based on discrete logarithm cryptography demonstrated higher efficiency compared to those based on integer factorization cryptography.
3. Based on homomorphic encryption: Freedman and his team [12] innovatively represented elements as roots of polynomials and encrypted the coefficients of these polynomials using Paillier homomorphic encryption technology, combined with zero-knowledge proofs, to implement a two-party PSI protocol resistant to malicious attacks. In 2016, Freedman et al. [13] further improved computational efficiency through the ElGamal encryption mechanism and reduced the protocol's computational complexity using Cuckoo Hash technology [7]. Abadi et al. [14] introduced a set representation method based on point-value pairs of d-degree polynomials, implemented through the Paillier encryption scheme, reducing the multiplication complexity from $O(d^2)$ to $O(d)$ [7]. Kissner and other researchers [15] adopted different polynomial representation methods, significantly reducing computational costs to be linearly proportional to the number of participants. Jarecki and others [16] used additive homomorphic encryption and zero-knowledge proofs to implement pseudo-random functions (PRF). Hazay and others [17] developed an additive homomorphic encryption scheme that supports threshold decryption for implementing multi-party semi-honest PSI protocols. Dou Jiawei and others [18] combined Paillier encryption to propose a PSI protocol based on the formula for calculating the area of triangles and rational number encoding.

Public key encryption-based PSI schemes typically feature fewer communication rounds and are suitable for environments with strong computational capabilities. However, in practice, communication bandwidth and time complexity often pose significant constraints [7].

2.1.2. Design Framework Based on Garbled Circuits

Garbled circuit technology can transform any function into a Boolean circuit, thereby securely computing the function. Early methods based on universal circuits, like the DPSZ scheme [19], demonstrated how to use arithmetic circuits to solve the set intersection problem: the circuit builder encrypts the circuit gates using a symmetric key, then creates a garbled circuit and sends it to the circuit evaluator. The evaluator decrypts specific paths in the garbled circuit to obtain the intersection results, while being unable to access other paths in the circuit. As the circuit depth increases, its construction complexity also increases. Additionally, PSI protocols based on this circuit design can also perform various symmetric function operations, such as calculating the threshold intersection,

the number of intersection elements, and their sum. For adversaries under semi-honest conditions, there are two types of garbled circuits: the Yao [20] protocol and the GMW [21] protocol. Pinkas et al. [6,22,23] and Chandran et al. based on hash storage structures and GMW circuits, implemented a more efficient OPRF circuit PSI scheme through private membership tests, reducing the number of comparisons and the depth of circuit equivalence comparisons. Meanwhile, Huang [24] and others created a semi-honest secure disordered circuit PSI scheme through the combination of Yao circuits, performing equivalence tests and specific sorting on adjacent elements. Despite these advantages, these methods still require additional key calculations and communication processes, such as key exchanges between participants.

2.1.3. Design Framework Based on Oblivious Transfer

Oblivious Transfer (OT) [25] is a cryptographic protocol that allows a sender to transmit information to a receiver without revealing any private information. In the OT protocol, the sender has two options, but the receiver can only obtain information about one of them without access to the other, and the sender does not know which option the receiver has chosen. OT is widely used in many secure areas due to its cryptographic robustness and privacy features. Its applications include secure protocol negotiation, secure online auction systems, and secure voting systems. In 2013, Dong [26] et al. proposed a new data structure—the garbled Bloom filter (GBF)—and based on the GBF and OT extension, they introduced a PSI protocol. This protocol utilized efficient symmetric encryption operations and could handle billions of elements. However, this protocol faced two issues: one is that the malicious sender might send incorrect shared information, and the other is that the input datasets are not independent. To address these problems, in 2016, Rindal and Rosulek [27] proposed a new randomized garbled Bloom filter using the "cut-and-choose" technique. They successfully implemented a two-party malicious model PSI protocol. Subsequently, Zhang [28] et al., based on this scheme, further proposed and implemented a multi-party PSI protocol, ensuring malicious security in the presence of two non-colluding servers, with computational and communication costs depending on the number of participants. Pinkas et al. [29] based on the OOS17-OT [30] protocol, built a maliciously secure PSI protocol. Rindal [31] et al., based on the semi-honest secure Schoppmann et al. [32] protocol and the maliciously secure Weng et al. [33] protocol, proposed, respectively, maliciously secure and semi-honest secure PSI protocols. Overall, PSI protocols based on oblivious transfer typically feature lower computational and communication overhead.

2.2. Unbalanced PSI

In recent years, many research articles [34–47] have focused on unbalanced PSI, primarily aiming to improve communication complexity and reduce the overall runtime of the protocols. However, none of these articles have specifically addressed the issue that, in unbalanced PSI scenarios, the party with the smaller dataset is often a mobile device with limited computational and storage capabilities. The complex cryptographic operations and filter storage requirements pose a significant burden on these devices. Therefore, the focus of this article is to alleviate the computational and storage burden on the party with the smaller dataset in unbalanced PSI scenarios while ensuring adequate security conditions are met.

2.3. Private Contact Discovery

As shown in Figure 1, mobile privacy contact discovery refers to when you install a communication application such as WhatsApp on your phone; the first thing this app does is check your contact list to see which of your contacts are also using their service. To achieve this functionality, the application could simply tell the service provider about the users in your contact list. The service provider can then inform you which of these users are also using their service. A typical example is WeChat, which recommends friend accounts based on the mobile phone's contact list. In this process, the WeChat server cross-matches

the mobile numbers in the user's contact list with its own WeChat account database, thus identifying all users in the contact list who have registered WeChat accounts, and provides friend account recommendations based on this information.

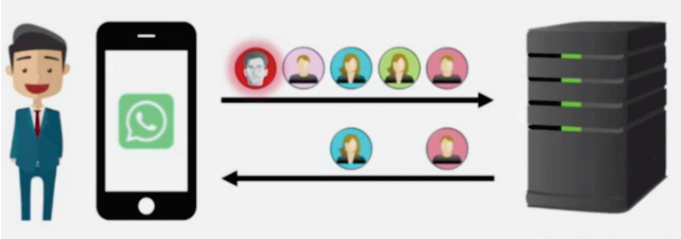

Figure 1. Private contact discovery.

This naturally involves privacy issues, as through our contacts, many aspects of our lives can be inferred, such as spending level and home address, and more seriously, it can lead to the leakage of our entire social graph. Furthermore, if one communicates with a highly influential person who holds many secrets, attackers could also potentially access these secrets through you.

The insecure solution to the privacy contact discovery problem allows users to send their contact set to the service provider, who then performs the intersection on behalf of the users. While this protects the privacy of the service provider, it exposes the users' private contacts to the service provider. In much research, current social media privacy contact discovery recommendation methods mainly include: (i) not offering privacy protection promises, (ii) seeking privacy protection through ambiguity (for example, using multiple signals for recommendations without explaining the reasons [48]), and (iii) using temporary thresholds to block the simplest attacks (typically, these thresholds are applied to the number of mutual friends between two users to decide whether to recommend one to the other). The only known attempt to rigorously address this issue in practice was the recent study by Signal [49], but it addressed a more limited problem than the one discussed in this paper—privately finding out whether contacts in a phonebook are Signal users. Theoretical research suggests adopting structured graph perturbation [50] and randomized recommendation methods [51,52] to achieve strict privacy guarantees of differential privacy [53]. Although differential privacy has begun to see practical application in other data mining applications [54], theoretical methods for privacy contact discovery have not yet been deployed, possibly because the trade-offs between privacy and utility they require are too harsh (i.e., they often recommend people who users are almost unlikely to know, significantly negatively impacting user experience quality). William Brendel [55] and others proposed a method using an auxiliary graph for deanonymization. Instead of modifying existing privacy contact discovery algorithms to protect privacy, they modify the graph used by the algorithm to create a candidate graph more resistant to practical brute force attacks, aiming to improve privacy with auxiliary graph information.

Some applications use a naive hashing protocol, where the client sends only the hash results of phone numbers to the service provider. Unfortunately, this technique is almost considered insecure. Because the entropy of phone numbers is low, naive hashing methods are susceptible to dictionary attacks. Generally speaking, PSI protocols are provably secure cryptographic protocols. They allow two parties to compute the intersection of their input sets without revealing any information other than the intersection. However, in mobile contact discovery scenarios, a common problem with most PSI protocols is that the online phase of the protocol, the computation of the intersection, has communication complexity linearly related to the size of the two input sets. In mobile contact discovery scenarios, the server-side database may contain millions or even billions of entries, while it is generally assumed that the client has about 1000 contacts. This makes the communication complexity of the protocol extremely high and impractical to apply.

3. Related Theories and Technologies

3.1. Multi-Party Secure Computation Security Model

The mathematical concept of Multi-Party Computation (MPC) involves several participants (such as $P_1, P_2, \ldots, P_n$), with each holding private input data (x_i). These participants collaboratively execute a computation of the function $f(x_1, x_2, \ldots, x_n)$ with the goal of ensuring that each participant can only access their own computational results, while being unable to ascertain the inputs and results of others. There are generally two security models employed in secure multi-party computation protocols [56]:

1. **Semi-honest model:** In this model, participants adhere to the protocol's execution rules but may attempt to gather other participants' inputs, outputs, and any accessible information during the execution of the protocol. This model assumes that the participants do not deviate from the established procedural rules but will use all available information to deduce the private data of others.
2. **Malicious adversary model:** Unlike the semi-honest model, the malicious adversary model accounts for the possibility that attackers may manipulate a subset of the participants to perform illicit actions, such as submitting incorrect input data or maliciously altering data to steal the private information of honest participants. Malicious adversaries might also disrupt the protocol by intentionally terminating its execution or by refusing to participate, thus preventing the protocol's completion.

The security model considered in this paper is the semi-honest security model.

3.2. Cuckoo Filter

Determining whether a particular element belongs to a given set is a common problem in computer science, with widespread applications in bioinformatics, machine learning, computer networks, the Internet of Things, and database systems [57]. Filter data structures such as Bloom filter and Cuckoo filter can approximately determine if an element is part of a specified set and have been extensively applied in network routing [58], file merging [59], spam detection [60], and distributed systems [61].

Filter data structures are used to approximately ascertain if an element belongs to a specific set. In essence, for a given set S and a query element x, the filter can approximately inform the query whether "x is in S". "Approximately" here implies that if x is actually not in S, the filter has a small error probability p of wrongly indicating that "x is in S"; however, if x is indeed in S, the filter will always correctly return that "x is in S". Filter data structures sacrifice some query accuracy to enhance space and time efficiency. Unlike data structures that require storing complete information of each element for precise queries, filter approximate the presence of an element solely through partial information such as hash values or "fingerprints". Based on this principle, existing filter data structures are mainly categorized into two types: one type uses bit arrays as in Bloom filter; the other type, exemplified by Cuckoo filter, is based on element "fingerprints".

The Cuckoo filter [62] is an advanced retrieval structure made up of multiple buckets, each capable of containing several bits. Compared to Bloom filter, Cuckoo filter offer the significant advantage of supporting deletion of elements and having higher space efficiency. With equal storage space, Cuckoo filter can achieve more accurate search results and shorter search times. When querying an element, the time complexity for Cuckoo filter is $O(1)$, meaning constant time complexity. This indicates that the execution time for query operations does not increase with the number of elements in the filter, an important performance feature of the Cuckoo filter design.

In this paper, Cuckoo filter are used to store data on database servers.

3.3. Paillier Homomorphic Encryption

Homomorphic encryption is an encryption technology that allows computations to be performed on encrypted data and to obtain encrypted results, which, when decrypted, are consistent with the results obtained by performing the same computations directly on the original data. This means that homomorphic encryption enables data to be processed and

analyzed without revealing any content. It is an important technology for protecting online privacy, allowing cloud computing services to perform complex data processing tasks on users' encrypted data without accessing the actual data.

Paillier homomorphic encryption is a public-key cryptosystem that specifically supports homomorphic addition operations on encrypted data. The applications of the Paillier encryption scheme are extensive, and it can be used to protect the privacy and security of data. For example, in distributed computing, the Paillier encryption scheme can be used to encrypt data and transmit it to various nodes for processing, ensuring the security and privacy of the data. Furthermore, the Paillier encryption scheme can also be used to implement homomorphic secret sharing, PSI, and other application scenarios. Overall, the Paillier encryption scheme is an efficient homomorphic encryption scheme with a wide range of application prospects. This paper uses the Paillier cryptosystem in its final scheme. The homomorphic properties utilized in this paper are as follows (the final scheme is based on these two features):

1. **Additive Homomorphism**: If $c_1 = \text{Enc}(m_1)$ and $c_2 = \text{Enc}(m_2)$, then $\text{Dec}(c_1 \cdot c_2 \mod n^2) = m_1 + m_2$. This allows for performing addition operations on ciphertexts without needing to decrypt them first.
2. **Scalar Multiplication Homomorphism**: If $c = \text{Enc}(m)$, then $\text{Dec}(c^k \mod n^2) = k \cdot m$. This means that it is possible to perform multiplication operations between a ciphertext and a plaintext scalar without decryption.

This paper will utilize homomorphic encryption technology to construct the third protocol of this paper: unbalanced PSI protocol based on dual-cloud assistance.

4. PSI Protocol Constructed Based on DH Key Exchange Mechanism

Before proposing the first unbalanced PSI protocol of this paper, we introduces Meadows' PSI protocol constructed based on the DH key exchange mechanism [8]. As shown in Figure 2, the specific process is as follows:

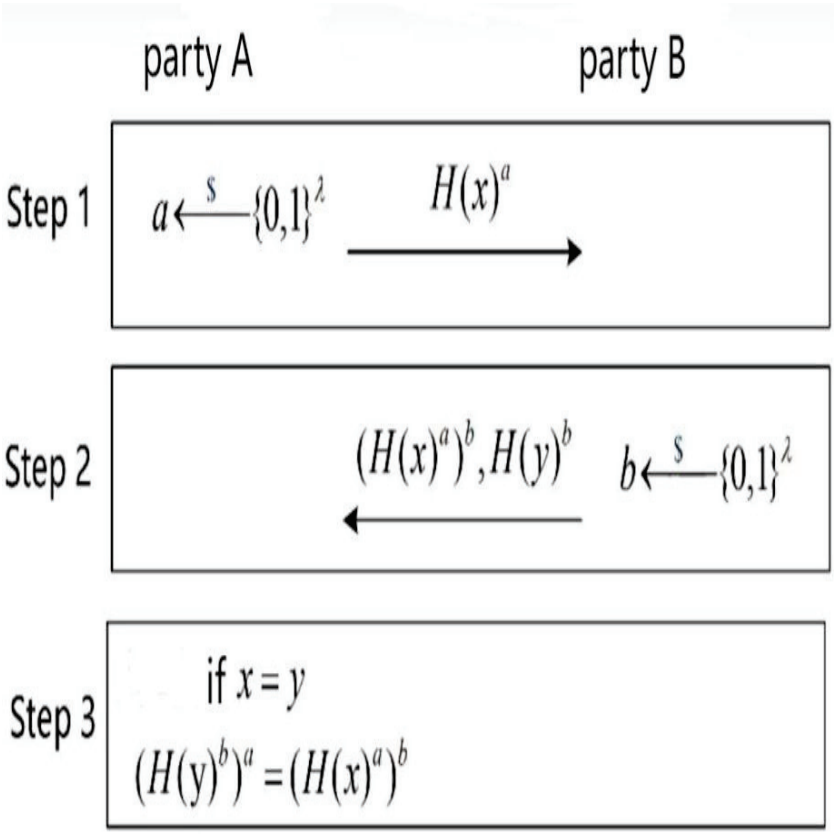

Figure 2. PSI protocol flow chart based on DH key exchange mechanism.

4.1. Protocol Process

4.1.1. Preprocessing Stage

1. Hash Processing: Each participant applies the same hash function to each element in their dataset to form a hash-processed dataset, ready for subsequent encryption and computation processes.

4.1.2. Exchange and Computation Stage

1. Exponentiation: Participant A takes each element from its dataset, use a to perform exponentiation operations (where a is A's private key) and forms a new set.
2. Data Exchange: Participant A sends the above-computed set to Participant B.
3. Auxiliary Dataset Construction: Upon receiving the dataset from A, Participant B uses b to perform exponentiation operations (where b is B's private key) to build an auxiliary dataset and sends the result set back to A.
4. Exponentiation: At the same time, Participant B also uses b to perform exponentiation operations of each element in its own dataset, which is also sent to A.

4.1.3. Intersection Identification Stage

1. Exponentiation and Comparison: After receiving two datasets from B, participant use a to perform exponentiation operations of the elements in the latter dataset received that has been powered by B. Then, A compares this result with another dataset received from B.
2. Intersection Determination: If an element after being powered a times matches an element in the auxiliary dataset sent by B, then that element belongs to the intersection of datasets A and B.

This protocol effectively protects the participants' data privacy through two exponential operations and hash processing, preventing data leakage during the exchange process.

4.2. Experimental Analysis

For this protocol, experiments were conducted and the runtime was recorded for various combinations of dataset sizes, as shown in Table 1, where the dataset cardinality refers to the number of elements in the dataset.

Table 1. Runtime of the PSI protocol Based on DH key exchange mechanism.

Cardinality of Dataset from Participant One	Cardinality of Dataset from Participant Two	Protocol Runtime (s)
2^{10}	2^{15}	1.7442
2^{10}	2^{17}	6.8024
2^{10}	2^{20}	55.2655
2^{10}	2^{25}	1849.2111
2^{15}	2^{15}	4.9248
2^{15}	2^{17}	10.0466
2^{15}	2^{20}	58.6051
2^{15}	2^{25}	1852.7098
2^{17}	2^{17}	20.0932
2^{17}	2^{20}	68.9472
2^{17}	2^{25}	1863.5443
2^{20}	2^{20}	165.4733
2^{20}	2^{25}	1964.6669

Through the experimental data, an important phenomenon can be observed. Table 1 shows the estimated runtime of the above protocol under different data volume levels. For example: Initially, when the number of elements in Participant One's dataset is 2^{10} and Participant Two's dataset is 2^{15}, the runtime of the protocol is 1.7442 s. In this case, there is a noticeable imbalance between the smaller side (Participant One) and the larger side

(Participant Two). Now, if we expand the number of elements in Participant One's dataset (originally the side with fewer elements) to 2^{15}, while keeping Participant Two's dataset size constant at 2^{15}, the runtime increases to 4.9248 s, approximately three times the original. This indicates that although the runtime increases when the datasets are balanced, the increase is limited. However, if we keep Participant One's dataset size at 2^{10} and increase Participant Two's dataset size to 2^{20} (the same scale of change), the runtime dramatically increases to 55.2655 s, about 31 times the initial condition. This phenomenon shows that in unbalanced dataset conditions, increasing the number of elements in the larger dataset significantly affects the efficiency of the protocol.

These results reveal the importance of dataset balance in maintaining efficiency during the implementation of this protocol. Unbalanced datasets not only lead to extended runtimes but can also cause low resource utilization and delays in processing. However, in practical applications, when two parties want to perform PSI, their sets are often unequal and with a significant gap. Therefore, the current situation requires the design of a new protocol to eliminate the impact of dataset size imbalance on protocol efficiency.

5. Unbalanced PSI Protocol Based on Cuckoo Filter

Although Meadows' PSI protocol [8] constructed based on the DH key exchange mechanism provides an effective way to compute the intersection of two datasets, especially under the premise of protecting participants' data privacy, this paper observes that its efficiency is significantly impacted when dataset sizes are extremely unbalanced. In particular, as shown in Table 1, the runtime increases significantly as the size of the larger dataset increases, reflecting the performance limitations of Meadows' PSI protocol when dealing with unbalanced datasets.

In order to overcome these limitations, the first protocol proposed in this paper adopts a different technical strategy, which effectively reduces the computational burden under unbalanced conditions by introducing Cuckoo filter. This not only optimizes the data processing process, but also improves the overall operational efficiency. In the new protocol, the increase in run time is not as dramatic as that in the Meadows [8] PSI protocol based on the DH key exchange, even with unbalanced dataset sizes, this allows for more efficient and balanced data processing. This improvement is particularly important for datasets of different sizes frequently encountered in practical applications.

Therefore, based on the above introduction, as shown in Figure 3, this paper first proposes a one-way PSI protocol based on the discrete logarithm problem difficulty and the correctness (high false positive rate) of Cuckoo filter. This protocol is divided into two phases, the specific details of which will be introduced later, and in the subsequent sections, the specific implementation details and optimization strategies of the protocol will be thoroughly explained.

5.1. Definition of Main Participants and Related Symbols

1. database server: represents the database server that holds all user data.
2. client: represents the mobile client who wants to perform private contact discovery services.
3. X and Y represent the dataset of the database server and the client, respectively.
4. α represents the private key of the database server in the Diffie–Hellman encryption algorithm.
5. β_j represents the random number generated by the client for the Diffie–Hellman encryption algorithm.
6. H represents the hash function negotiated by the client and database server for use.
7. CF represents Cuckoo Filter, $CF.insert$ represents the operation of adding an element to the Cuckoo filter, $CF.check$ represents the operation of checking whether a specific element exists in the filter.
8. X_i represents the i-th element of set X. Similarly, Y_i, C_i, etc., also represent similar meanings.

9. $C = \{c_1, c_2, \ldots, c_n\}$ represents the set containing n_2 ciphertexts sent by the client to the database server.
10. C' represents the set containing n_2 ciphertexts sent by the database server to the client.
11. $item_j$ represents the result obtained through a series of exchange and decryption operations, used to retrieve the filter to obtain the intersection.

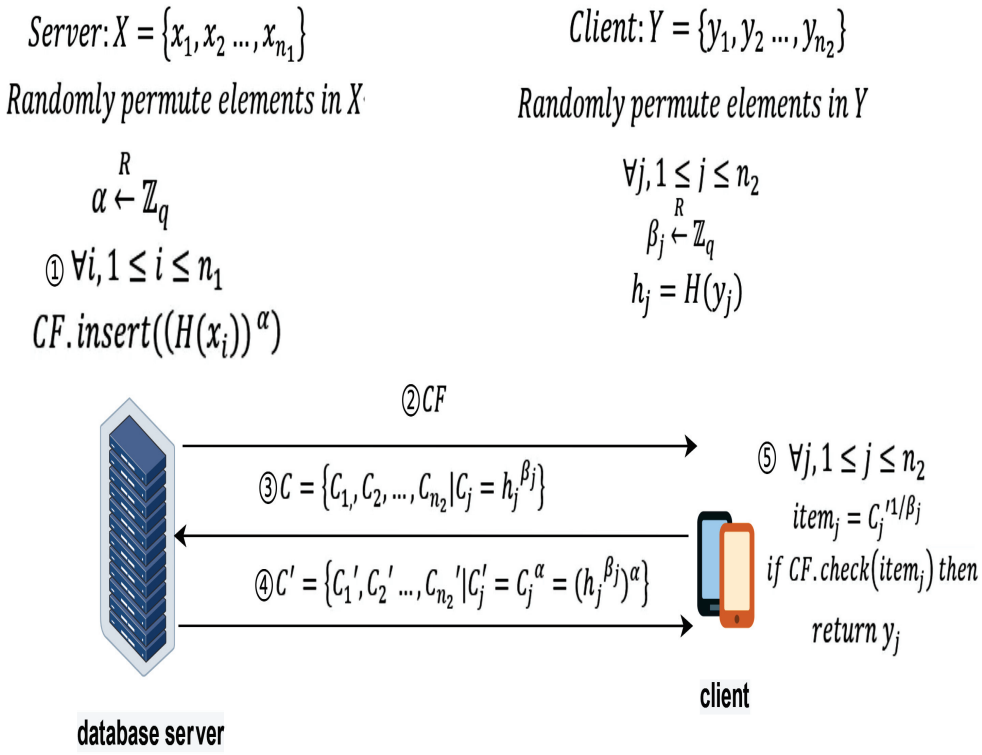

Figure 3. Unbalanced PSI protocol based on Cuckoo filter.

5.2. Protocol Process

The protocol is divided into two phases: the preprocessing phase and the intersection phase, with specific details as follows.

5.2.1. Preprocessing

In the preprocessing phase, the client and server need to perform a series of preparatory work to ensure the security and efficiency of subsequent interactions. The specific steps are as follows:

1. Security parameter negotiation: The client and database server agree on the large prime number q used in the DH encryption algorithm and the hash function H used.
2. Database server generates private key: The database server generates its own private key α, used for the Diffie–Hellman (DH) encryption algorithm.
3. Data scrambling: The client and database server scramble their own datasets X and Y for randomization, enhancing data privacy and security.
4. Client data preprocessing: The client calculates $h_j = H(y_j)$ and generates n_2 random numbers β_j, used for the Diffie–Hellman (DH) encryption algorithm.
5. Creation of Cuckoo filter: The database server generate a Cuckoo filter CF by using the operation $CF.insert((H(x_i))^{\alpha})$, and sends the filter CF to the client for PSI queries with privacy protection.

5.2.2. Intersection

In the intersection phase, the client and database server perform a series of carefully designed encryption and decryption operations to blind the client's elements securely and compute the intersection of the two sets. The specific operations are as follows:

1. Element blinding and interactive encryption operations: The client and the database server interact through a series of asymmetric encryption and decryption operations to blind the client's elements. Specifically, the client calculates $C_j = h_j^{\beta_j}$ and sends C to the database server. The database server uses its private key α to compute $C'_j = C_j^{\alpha}$ and sends C' back to the client.
2. Intersection computation: After receiving C', the client checks whether they belong to the filter CF through the check operation $CF.check$, thereby calculating the intersection of the sets. Specifically, after receiving C' sent by the database server, the client computes $item_j = C_j'^{1/\beta_j}$ and uses the result to query the filter CF to obtain the intersection element y_j.

5.3. Correctness Analysis

If $x_i = y_j$, then $H(x_i) = H(y_j)$, then $item_j = C_j'^{1/\beta_j} = C_j^{\alpha * 1/\beta_j} = \left(h_j^{\beta_j}\right)^{1/\beta_j * \alpha} = h_j^{\alpha} = H(x_i)^{\alpha} = H(y_j)^{\alpha}$.

Thus, through this scheme, the client can accurately obtain the intersection elements of both parties.

5.4. Security Analysis

This section will analyze the security of the protocol in detail, mainly its ability to protect the privacy of both parties.

Firstly, considering that the protocol utilizes the Diffie–Hellman (DH) key exchange mechanism to blind the client's elements, this process's security is based on the difficulty of solving the One-More-Gap-Diffie–Hellman (OMGDH) problem. Since the DH mechanism ensures that, even in public communication channels, unauthorized parties cannot decipher the exchanged secret information, the client's data are protected during transmission to the server. The server uses a private key to process the received data and returns the results to the client; this process likewise ensures the security and privacy of the data server.

Secondly, the protocol's use of Cuckoo filter means that although it efficiently supports insertion and query operations, its false positive characteristics mean that even if some non-intersecting elements are mistakenly identified as belonging to the intersection, it does not reveal the exact set membership information. This feature provides additional privacy protection to some extent, as even in the event of a false positive error, attackers cannot determine whether a specific element truly exists in the other party's set.

Furthermore, through the interactive computations between the client and the server, the protocol ensures that only elements common to both parties can be accurately identified. The client checks the data returned by the server against its own dataset to ultimately determine the intersection.

In summary, based on the blinding process using the Diffie–Hellman mechanism and the use of Cuckoo filter, this protocol can accurately calculate the intersection of two sets while protecting the participants' privacy.

5.5. Experimental Analysis

For this protocol, experiments were conducted, and the runtime was recorded for various combinations of data volumes, as shown in Table 2. The table also compares the runtime of Meadows' PSI protocol constructed based on the DH key exchange mechanism [8]. Since preprocessing can be completed offline, the runtime of the unbalanced PSI protocol based on Cuckoo filter refers to the total time of the outsourcing process and the intersection process. The original protocol refers to the PSI protocol constructed based on the DH key exchange mechanism, and the new protocol refers to the unbalanced PSI protocol based on Cuckoo filter.

Table 2. Comparison of PSI protocol runtime based on DH key exchange mechanism and Cuckoo filter.

Cardinality of Dataset from Participant One	Cardinality of Dataset from Participant Two	Original Protocol Runtime (s)	New Protocol Runtime (s)
2^{10}	2^{15}	1.7442	0.1539
2^{10}	2^{17}	6.8024	0.1569
2^{10}	2^{20}	55.2655	0.1616
2^{10}	2^{25}	1849.2111	0.1693
2^{15}	2^{15}	4.9247	4.9239
2^{15}	2^{17}	10.0465	5.0232
2^{15}	2^{20}	58.6050	5.1709
2^{15}	2^{25}	1852.7097	5.4172
2^{17}	2^{17}	20.0931	20.0930
2^{17}	2^{20}	68.9471	20.6841
2^{17}	2^{25}	1863.5442	21.6690
2^{20}	2^{20}	165.4732	165.4731
2^{20}	2^{25}	1964.6668	173.3531

Through the experimental data, this paper can observe several key phenomena. First, when the cardinality of the smaller dataset (number of elements) remains constant while the number of elements in the larger dataset increases rapidly, it is observed that the runtime of the protocol does not change much, remaining consistent. This indicates that although the size of the large dataset increases dramatically, the efficiency of the protocol is not significantly affected, thereby proving that the design of this protocol can effectively mitigate the negative impact of dataset size imbalance on protocol efficiency. Especially, the overall runtime of the protocol is more related to the cardinality of the smaller dataset and has very low relevance to the cardinality of the larger dataset.

At the same time, this experiment also found that under balanced dataset conditions, the runtime of the PSI protocol based on the DH key exchange mechanism and the unbalanced PSI protocol based on Cuckoo filter does not differ significantly. This indicates that when the sizes of the sets are similar, both protocols can exhibit comparable performance, providing an efficient data intersection solution.

5.6. Summary of This Chapter

In this chapter, two types of PSI protocols are discussed in detail: the PSI protocol constructed based on the DH key exchange mechanism [8] and the unbalanced PSI protocol based on Cuckoo filter. By comparing the runtime of these two protocols under different data volume conditions, this chapter has obtained a series of important observations and conclusions.

First, the PSI protocol based on DH key exchange shows a performance decline when dealing with imbalanced datasets. Especially when the size of the smaller dataset remains constant while the size of the larger dataset increases significantly, the runtime of this protocol increases dramatically, showing efficiency issues when processing large datasets.

In contrast, the unbalanced PSI protocol based on Cuckoo filter exhibits more stable performance under various conditions of dataset size imbalance. Even when the size of the large dataset increases significantly, the change in runtime for this protocol is not substantial, proving its superiority in dealing with unbalanced datasets. Additionally, when the datasets are nearly balanced, the performance difference between the two protocols is not significant, indicating that both protocols can work effectively under balanced dataset conditions.

Therefore, facing settings with dataset imbalance, the unbalanced PSI protocol based on Cuckoo filter appears more efficient. It not only maintains a relatively stable runtime in situations where the dataset sizes are extremely unbalanced but also exhibits performance comparable to the PSI protocol based on DH key exchange even when the

dataset sizes are similar. This robustness makes the unbalanced PSI protocol based on Cuckoo filter a more ideal choice when facing the common problem of dataset imbalance in practical applications.

6. Unbalanced PSI Protocol Based on Single-Cloud Assistance

The previous chapter has proven that the unbalanced PSI protocol based on Cuckoo filter is more suitable for practical scenarios; especially under unbalanced conditions, this protocol effectively resolves the performance limitations of the PSI protocol constructed using the DH key exchange mechanism in handling unbalanced datasets. However, there is still room for improvement in this protocol. It is observed that in this protocol, clients need to store filter and perform complex cryptographic operations, which can be a significant burden for mobile devices with limited computing power and storage space. A series of encryption operations and storing filter received from the other party becomes a heavy load. To address this issue, it is considered to transfer most of the client's computational and storage tasks to cloud servers. By delegating tasks to cloud servers, clients can significantly reduce computational and storage pressure, especially for mobile devices with limited capabilities.

This section will introduce cloud computing technology, which allows clients with limited computing power and storage space to outsource their private data and request cloud platforms to perform related computations. Currently, whether for individual users or large enterprises, entrusting data storage and computation tasks to cloud services has become a common practice. Based on the introduction above, as shown in Figure 4, this chapter proposes a second unbalanced PSI scheme based on the unbalanced PSI protocol using Cuckoo filter.

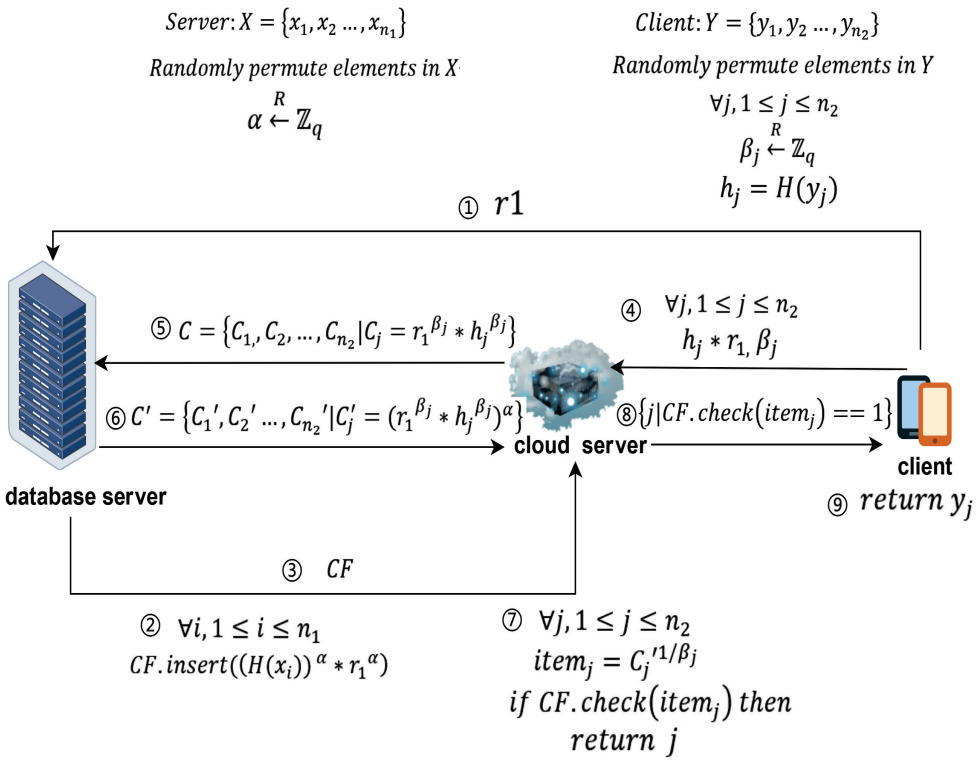

Figure 4. Unbalanced PSI protocol based on single-cloud assistance.

6.1. Definition of Main Participants and Related Symbols

1. database server: represents the database server that holds all user data.
2. client: represents a mobile client that wants to discover private contacts.
3. cloud server: represents an auxiliary server that assists the client in performing intersection operations, undertaking most of the computational and storage pressures.

4. X and Y, respectively, represent the dataset of the database server and the client dataset.
5. α represents the private key of the database server in the Diffie–Hellman encryption algorithm.
6. r_1 represents the random number generated by the client, used to blind the data.
7. β_j represents the random number generated by the client for the Diffie–Hellman encryption algorithm.
8. H represents the hash function negotiated for use by the client and database server.
9. CF represents the Cuckoo Filter, $CF.insert$ represents the operation to add an element to the Cuckoo filter, $CF.check$ represents the operation to check if a specified element exists in the filter.
10. X_i represents the i-th element of the set X. Similarly, Y_i, C_i, etc., also represent similar meanings.
11. $C = \{c_1, c_2, \ldots, c_n\}$ represents the set of n_2 ciphertexts sent by the client to the database server.
12. C' represents the set of n_2 ciphertexts sent by the database server to the client.
13. $item_j$ represents the result obtained through a series of exchange and decryption operations, used to retrieve the filter to obtain the intersection.

6.2. Protocol Process

6.2.1. Preprocessing

1. Security parameter negotiation: Each role discusses the necessary security parameters—all parties share the large prime q used in the DH cryptographic algorithm. The client and database server negotiate to generate r_1 and the hash function H.
2. Database server generates a private key: The database server generates its own private key α, for use in the Diffie–Hellman encryption algorithm.
3. Data scrambling: The client and database server each scramble their own datasets X and Y.
4. Client data preprocessing: The client calculates $h_j = H(y_j)$, generates n_2 random numbers β_j, and calculates $h_j * r_1$.

6.2.2. Outsourcing

1. Database server sends data to the cloud server: The database server uses its private key α to perform the operation $CF.insert((H(x_i))^\alpha * r_1^\alpha)$, creates a Cuckoo filter CF, and sends it to the cloud server.
2. Client sends data to the cloud server: The client sends the random numbers β_j and $h_j * r_1$ to the cloud server. After receiving the data sent by the client, the cloud server calculates $C_j = r_1^{\beta_j} * h_j^{\beta_j}$. At this point, the cloud server has saved the client's blinded data.

6.2.3. Intersection

1. Cloud server sends data: The client cloud server sends the blinded data C_j to the database server.
2. Database server processes data: Upon receiving C_j, the database server uses its private key α to calculate $C'_j = C_j^\alpha$, and sends the result back to the cloud server.
3. Cloud server processes data: After receiving C'_j from the database server, the cloud server calculates $item_j = C_j^{1/\beta_j}$ and uses the result to search CF. If $item_j$ exists in CF, it returns the index j of $item_j$ and sends j to the client.
4. Obtaining the intersection: The client obtains the intersection element y_j through the received index j.

6.3. Correctness Analysis

If $x_i = y_j$, then $H(x_i) = H(y_j)$, so $item_j = C_j'^{1/\beta_j} = C_j^{\alpha * 1/\beta_j} = \left(r_1{}^{\beta_j} * h_j^{\beta_j}\right)^{1/\beta_j * \alpha} = r_1{}^{\alpha} * h_j^{\alpha} = r_1{}^{\alpha} * H(x_i)^{\alpha} = r_1{}^{\alpha} * H(y_j)^{\alpha}$.

Thus, through this scheme, the client can accurately obtain the intersection elements of both parties.

6.4. Security Analysis

In the design of this protocol, the primary security objective is to ensure that, even in a partially trusted cloud environment, neither the client's data nor the database server's data can be accessed or inferred by unauthorized entities. Specifically, since other participating parties are unaware of the database server's private key α, they cannot deduce the data held by the database server. Similarly, since other parties do not know the client's private random number r_1, they cannot deduce the client's data.

However, this scheme has inherent security risks, primarily because it does not withstand collusion attacks. If the data server and the cloud server collude, they can jointly deduce the client's data. This is possible because the cloud server possesses the blinded data $h_j \times r_1$, and if the data server leaks the private key r_1 to the cloud server, then both the cloud server and the database server could deduce the client's original data h_j. Collusion attacks are a security threat where two or more distinct entities (for example, users, systems, or service providers) secretly cooperate to undermine or circumvent security mechanisms and privacy measures. In cloud computing environments, cloud service providers and cloud users may collude to steal or infer other users' sensitive data stored on the cloud.

6.5. Experimental Analysis

6.5.1. Data Storage Volume

In the research of this paper, the experimental analysis of the unbalanced PSI protocol based on single-cloud assistance revealed a key issue: when the client needs to receive a Cuckoo filter from the database server, this poses a significant challenge for mobile clients with limited storage capacity. This challenge is magnified when facing large datasets.

To understand this issue deeply, a series of experiments were conducted to measure the volume of Cuckoo filter needed by the client under different data sizes. The input data size for the experiments was provided by the database server, reflecting the various data volumes that might be encountered in actual application scenarios. As shown in Table 3, the paper meticulously recorded the specific sizes of Cuckoo filter under different input data volumes, revealing the intrinsic relationship between data volume and filter size. Through experiments, it was discovered that as the data volume in the database server increased, the storage burden on the client under the original protocol also increased accordingly, with the size of the Cuckoo filter directly impacted by the input data volume. In applications highly dependent on contact discovery services and containing extensive user information, this storage pressure is particularly evident. For example, for applications containing hundreds of millions of data entries, the volume of the Cuckoo filter could reach an overwhelming 14.850 GB, a significant challenge for most mobile devices.

However, the proposed protocol based on cloud assistance significantly alleviates this pressure. In the protocol, the Cuckoo filter is stored on the cloud server, thereby avoiding direct data transmission to the client. This approach not only effectively reduces the client's storage burden but also maintains the system's efficient operation, especially when handling large-scale data. By making such system design adjustments, the paper not only ensures the protection of data privacy but also significantly improves the feasibility and practicality of contact discovery services in actual applications.

Table 3. Size of Cuckoo filter at different data volumes.

Data Set Count	Size of Cuckoo Filter (MB)
2^{15}	0.535
2^{17}	2.363
2^{20}	21.678
2^{22}	93.645
2^{23}	194.436
2^{24}	403.201
2^{27}	3571.206
2^{28}	7372.835
2^{29}	15,206.421

In summary, the experimental results of this chapter emphasize the importance of optimizing storage strategies when dealing with large data scenarios. By outsourcing some storage tasks to the cloud server, the solution proposed in this chapter offers a viable approach for mobile clients needing to perform private contact discovery services, addressing the growing demands for data storage.

Therefore, based on the above analysis and experimental results, it is clear that when the data volume in the database server is excessively large, in other words, when the number of users reaches a certain level, the feasibility of a simple unassisted unbalanced PSI protocol based on Cuckoo filter significantly decreases, especially in applications like private contact discovery. This is because the unbalanced PSI protocol based on Cuckoo filter requires clients to directly receive and process massive Cuckoo filter, which poses a significant challenge for clients with limited storage resources, particularly mobile devices. Client devices often do not have enough storage space to accommodate these large-volume filter data, let alone process these data to complete PSI operations.

In this context, the introduction of a cloud server scheme shows its unique advantages. By transferring the storage of the filter to the cloud server, the burden on the client is greatly reduced. By this means, even in situations with a massive number of users and large data volumes, the scheme can still maintain efficient operations and ensure the smooth completion of PSI operations.

In summary, through experimental and theoretical analysis, this section concludes that in scenarios with large-scale users and massive data volumes, the introduction of a cloud server scheme is more feasible and efficient than the unbalanced PSI protocol based on Cuckoo filter.

6.5.2. Protocol Running Time

For this protocol, as shown in Table 4, the paper conducted experiments and recorded the running time of the protocol under various data volume combinations. Because pre-processing can be completed offline, the running time of the protocol refers to the total time of the outsourcing process and the intersection process. Table 4 also compares the running times of the unbalanced PSI protocol based on Cuckoo filter and the unbalanced PSI protocol based on single-cloud assistance. Here, Protocol 1 refers to the unbalanced PSI protocol based on Cuckoo filter, and Protocol 2 refers to unbalanced PSI protocol based on single-cloud assistance.

From the experimental analysis, the following conclusions can be drawn: In cases of smaller data volumes, the performance differences between the two protocols are not significant. However, as the data volume increases, the running time differences between different protocols gradually become apparent. This is because, at certain specific levels, the proportion of communication time is relatively high when the data volume is small, significantly impacting the results. For larger data volumes, where computation time dominates, Protocol 2, by placing computational tasks on the more powerful cloud server, gradually widens the running time difference from Protocol 1. Overall, the use of cloud

resources in the unbalanced PSI protocol based on single-cloud assistance significantly reduces running times, especially when dealing with large-scale datasets.

Table 4. Running times of Protocol 1 and Protocol 2 under different data volume combinations.

Client Dataset Size	Database Server Dataset Size	Protocol 1 Running Time (s)	Protocol 2 Running Time (s)
2^{10}	2^{15}	0.1539	0.1543
2^{10}	2^{17}	0.1569	0.1573
2^{10}	2^{20}	0.1616	0.1611
2^{10}	2^{25}	0.1693	0.1683
2^{15}	2^{15}	4.9239	3.8223
2^{15}	2^{17}	5.0232	3.9145
2^{15}	2^{20}	5.1709	4.0267
2^{15}	2^{25}	5.4172	4.2233
2^{17}	2^{17}	20.0930	15.6768
2^{17}	2^{20}	20.6841	16.1281
2^{17}	2^{25}	21.6690	16.8939
2^{20}	2^{20}	165.4731	129.0516
2^{20}	2^{25}	173.3531	135.2534

6.6. Summary of This Chapter

This chapter introduces an unbalanced PSI protocol based on single-cloud assistance, which utilizes cloud computing to reduce the computing and storage pressure of the client compared with previous protocols. Additionally, in the absence of collusion between the data server and the cloud server, the protocol effectively protects data from unauthorized access, ensuring the confidentiality of the data and the privacy of the client, making it highly suitable for scenarios where the cloud server is fully trusted.

However, it cannot be denied that although this scheme significantly reduces the computational and storage burden on the client, its security against collusion attacks is insufficient. When the possibility of collusion between the cloud server and data server cannot be completely ruled out, the protocol faces security risks and will require further security enhancement measures. Therefore, the next chapter will introduce a more secure solution to address the security deficiencies of the current scheme, ensuring the security and privacy of client data and database server data in environments where not all parties are fully trustworthy. In other words, the new scheme can resist collusion attacks.

7. Unbalanced PSI Protocol Based on Dual-Cloud Assistance

The previous single-server solution, which efficiently delegated computationally intensive encryption operations such as exponentiation and storage-intensive Cuckoo filter to the cloud server, has indeed alleviated the computational and storage burdens on the client to a certain extent. This is particularly advantageous for clients with limited computing and storage capabilities, allowing them to operate beyond their hardware constraints. However, security analysis reveals that the unbalanced PSI protocol based on single-cloud assistance has inherent security risks, specifically when collusion between the cloud and data servers is possible, thus compromising its adequacy in protecting client data privacy.

As shown in Figure 5, to preserve the advantages of the previous scheme—namely reducing computational and storage pressures on the client—while addressing these security issues, this chapter proposes a new solution. This design aims to enhance the security during data processing, especially against potential collusion attacks.

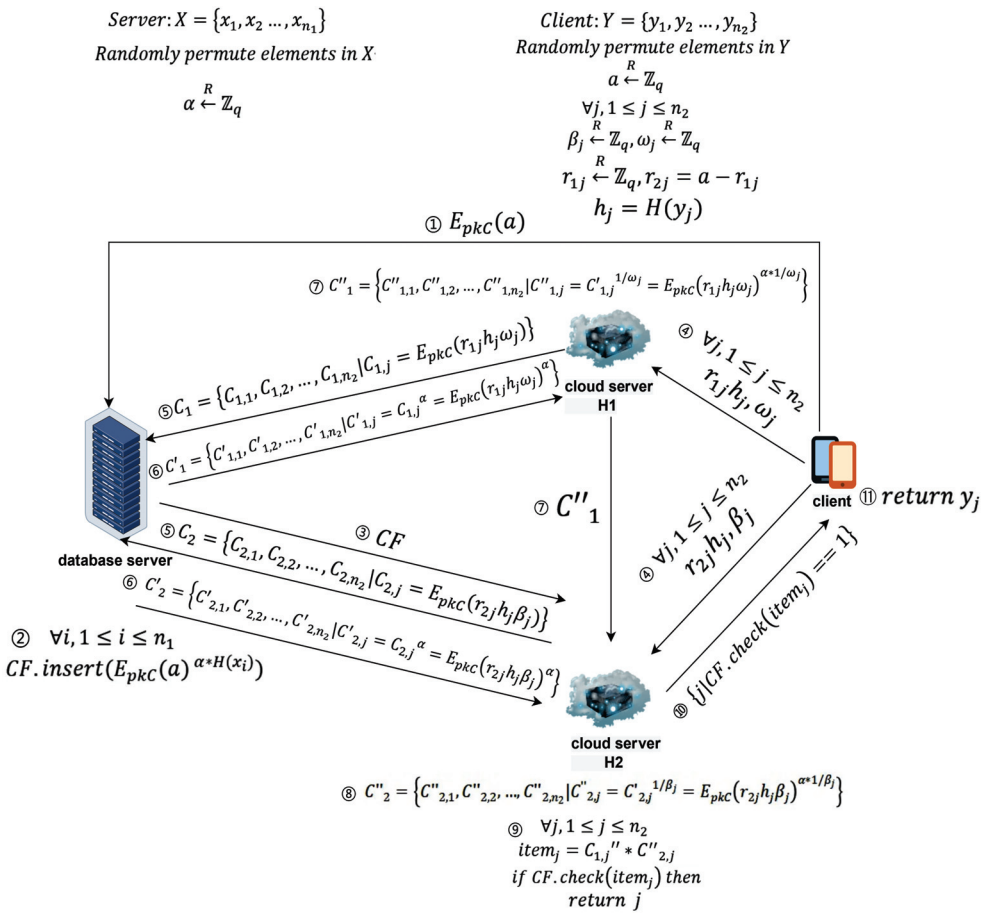

Figure 5. Unbalanced PSI protocol based on dual-cloud assistance.

7.1. Definition of Main Participants and Related Symbols

1. database server: represents the database server that holds all user data.
2. client: represents the mobile client that wishes to perform private contact discovery services.
3. cloud server H_1: acts as an auxiliary server for the client, handling the majority of computation and storage pressures.
4. cloud server H_2: another auxiliary server handling substantial computational and storage demands.
5. X and Y: represent the dataset of the database server and the client, respectively.
6. α: represents the private key of the database server used in the Diffie–Hellman encryption algorithm.
7. H: the hash function agreed upon by the client and the database server for use.
8. CF: represents the Cuckoo Filter, where $CF.insert$ denotes the operation to add elements, and $CF.check$ checks for the presence of specific elements.
9. ω_j: random exponentials generated by the client for cloud server H_1, β_j for cloud server H_2.
10. a: a secret value held by the client.h
11. $r_{1,j}$: random numbers used by the client for sending obfuscated data to cloud server H_1, and $r_{2,j}$ for H_2 where $r_{1,j} + r_{2,j} = a$.
12. C_1: the ciphertext collection sent from cloud server H_1 to the database server, and C_2 from H_2; $C_{1,j}$ and $C_{2,j}$ are specific elements within these collections.
13. C'_1 and C'_2: processed ciphertext collections returned to H_1 and H_2 from the database server; $C'_{1,j}$ and $C'_{2,j}$ are specific elements within these collections.
14. C''_1 and C''_2: final processed ciphertext collections at H_1 and H_2 after receiving data from the database server; $C''_{1,j}$ and $C''_{2,j}$ are specific elements within these collections.

15. $item_j$: represents the result of multiplying $C''_{1,j}$ and $C''_{2,j}$ used to query the filter.
16. j: represents the index used by the client to obtain the intersection.

7.2. Protocol Process

7.2.1. Preprocessing

1. Discuss security parameters: Each party discusses the necessary security parameters—the large prime q used in DH encryption and the client's public key pk_c required for the Paillier encryption system. The client and the database server negotiate the creation of hash function H.
2. Client sends $E_{pk_c}(a)$: the client generates its private secret number a and sends $E_{pk_c}(a)$ to the database server.
3. Database server generates private key: the database server creates its private key α, used for the DH encryption algorithm.
4. Data scrambling: the client and the database server each shuffle their respective datasets.
5. Client calculates hashes and generates random numbers: the client computes $h_j = H(y_j)$ and generates n_2 random numbers β_j, ω_j, r_{1j}, r_{2j}, and computes $r_{1j}h_j$, $r_{2j}h_j$ where $r_{1j} + r_{2j} = a$.

7.2.2. Outsourcing

1. Client sends data to cloud servers: the client sends $r_{1j}h_j$, ω_j to cloud server H_1, and $r_{2j}h_j$, β_j to cloud server H_2. H_1 computes $C_{1,j} = E_{pk_c}(r_{1j}h_j\omega_j)$, and H_2 computes $C_{2,j} = E_{pk_c}(r_{2j}h_j\beta_j)$. At this point, H_1 and H_2 hold the client's obfuscated data.
2. Database server sends data to cloud servers: Using $E_{pk_c}(a)$, the database server performs the filter insertion operation CF.insert$\left(E_{pk_c}(a)^{\alpha * H(x_i)}\right)$ to generate a Cuckoo filter and sends it to cloud server H_2. H_2 stores the filter sent by the database server.

7.2.3. Intersection

1. H_1 and H_2 send data: H_1 and H_2 each send their respective collections C_1 and C_2 to the database server.
2. Database server processes data: Upon receiving the data, the database server uses its private key α to compute $C'_{1,j} = C^{\alpha}_{1,j} = E_{pk_c}(r_{1j}h_j\omega_j)^{\alpha}$ and sends the results back to H_1. It also processes $C'_{2,j} = C^{\alpha}_{2,j} = E_{pk_c}(r_{2j}h_j\beta_j)^{\alpha}$ and sends the results back to H_2.
3. H_1 processes data: after receiving data from the database server, H_1 uses the random number ω_j to calculate $C''_{1,j} = C'^{1/\omega_j}_{1,j} = E_{pk_c}(r_{1j}h_j\omega_j)^{\alpha * 1/\omega_j}$ and sends the results to H_2.
4. H_2 processes data: Upon receiving data from H_1 and the database server, H_2 calculates $C''_{2,j} = C'^{1/\beta_j}_{2,j} = E_{pk_c}(r_{2j}h_j\beta_j)^{\alpha * 1/\beta_j}$. H_2 checks if $item_j = C''_{1,j} * C''_{2,j}$ exists in CF. If it does, it returns the index j of $item_j$ and sends it to the client.
5. Obtaining the intersection: the client receives the index j and retrieves the intersecting element y_j.

7.3. Correctness Analysis

If $x_i = y_j$, then $H(x_i) = H(y_j)$, which implies that $C''_{1,j} * C''_{2,j} = C'^{1/\omega_j}_{1,j} * C'^{1/\beta_j}_{2,j} = E_{pk_c}(r_{1j}h_j\omega_j)^{\alpha * 1/\omega_j} * E_{pk_c}(r_{2j}h_j\beta_j)^{\alpha * 1/\beta_j} = E_{pk_c}(r_{1j}h_j\alpha) * E_{pk_c}(r_{2j}h_j\alpha) = E_{pk_c}[\alpha h_j(r_{1j} + r_{2j})] = E_{pk_c}(\alpha h_j a) = E_{pk_c}(a)^{\{\alpha * H(x_i)\}}$.

Thus, the client can accurately obtain the intersection elements from both parties.

7.4. Security Analysis

Firstly, we consider the security of a small health app developer's data in the set. In considering security against collusion attacks, it is generally assumed that there is an

adversary who possesses the perspective and information of all participating parties except for the protected entity. This means the adversary can access, control, or receive information and resources from all participants except for the protected party. In this scenario, the adversary attempts to compromise the system's security or privacy by aggregating these insights, such as revealing sensitive data of the protected party. If, in this context, the adversary still cannot learn or infer the protected party's data, then it is proven that the data and privacy of the protected party are sufficiently secured against collusion attacks.

This section defines a game where the security objective is to maintain confidentiality of the data within the set under semi-honest and collusion conditions. The game for securing the client's dataset is as follows:

1. The client runs the preprocessing algorithm, sharing the cryptographic hash function H and the large prime q used in the protocol with the adversary.
2. The client simulates the outsourcing algorithm and sends their (encrypted) input to the adversary.
3. The client and the adversary simulate the intersection algorithm and discard any output.
4. The adversary is asked to output a guess $\hat{y}$ of the client's input y.

The game is analogized to a deterministic one-way function, such as a public key encryption scheme. Let S be the simulated messages of the client during the game. Let a one-way function adversary A' be given the information (public key) pk and function (ciphertext) c (encrypted y). The advantage of the adversary Adv_A is defined as the difference between the successful guesses of A and A'. If this advantage is negligible in the security parameter λ, then the outsourced PSI is considered secure. That is, let $Adv_A = \Pr[A(S) = y] - \Pr[A'(pk, c) = y]$. If $Adv_A < \frac{1}{\text{poly}(\lambda)}$, then the protocol is said to be secure.

Specifically, after the steps mentioned above, H_1, H_2, and the database server have a complete view of the process. However, under the two-server architecture, as illustrated in Figure 5:

1. In step four of Figure 5, since r_{1j} and r_{2j} are unknown to the adversary, h_j cannot be derived. The adversary can only attempt exhaustive guessing, thus making Adv_A negligible.
2. In subsequent steps, as A does not know the client's private key for the Paillier encryption system, it is impractical to decrypt the ciphertexts, making it even more challenging to derive h_j. For instance, $item_j = C''_{1,j} * C''_{2,j} = {C'_{1,j}}^{1/\omega_j} * {C'_{2,j}}^{1/\beta_j} = E_{pk_c}(r_{1j} h_j \omega_j)^{\alpha * 1/\omega_j} * E_{pk_c}(r_{2j} h_j \beta_j)^{\alpha * 1/\beta_j}$, and since the private key used in Paillier's system by the client is unknown, decrypting this compound is complex and hence h_j remains secure.

From the analysis above, it is evident that the advantage of Adv_A is negligible. Therefore, if both cloud servers collude with the database server, they cannot deduce the client's original data.

Next, consider the security of the data in the database server's set. Obviously, apart from the database server itself, none of the parties know the database server's private key α, hence even if both cloud servers colluded with the client, they cannot derive the original data from CF.

It is particularly noted that due to the prevalence of attacks on hash functions, further security enhancements are recommended by protecting the hashed data as the raw data.

In conclusion, the dual-server scheme successfully resists collusion attacks under semi-honest conditions. By thoroughly integrating considerations for security and privacy into the protocol design, both the client's and the database server's data are assured of robust protection. This solution not only provides an effective mechanism for PSI but also demonstrates resilience against potential collusion threats.

7.5. Experimental Analysis

7.5.1. Data Computation Volume

When evaluating any protocol's performance, the computational load borne by the client is undoubtedly a critical factor. As all three protocols have been introduced, this section focuses particularly on the client's computational volume, to accurately gauge and compare the efficiency of the three distinct protocols in operation. Specifically, this section will conduct a detailed analysis and comparison of the primary computational tasks that the client must execute across these protocols, to fully assess each protocol's demand on the client's computational resources. This analysis will primarily focus on the types of operations involved, aiming to clarify which protocol demonstrates relative advantages in reducing the client's computational burden, thus providing a solid basis for selecting the most appropriate protocol. Below is an analysis of the main types of operations involved in each protocol, focusing primarily on the outsourcing and intersection processes, as the preprocessing can be completed offline.

1. **Unbalanced PSI protocol based on Cuckoo filter:** two rounds of modular exponentiation operations and filter retrieval.
2. **Unbalanced PSI protocol based on single-cloud assistance**: a single round of multiplication operations and outputting y_j based on index j.
3. **Unbalanced PSI protocol based on dual-cloud assistance**: two rounds of multiplication operations and outputting y_j based on index j.

An analysis of the single-instance time consumption for these four operations offers a practical insight into the computational volume differences:

1. **Modular exponentiation operation**: Representing computation-intensive operations, modular exponentiation becomes particularly time-consuming. On a standard hardware setup, the time required for a single modular exponentiation operation depends primarily on the size of the numbers involved and the efficiency of the algorithm.
2. **Multiplication operation**: Compared to modular exponentiation, multiplication operations execute much faster on modern computing systems, even when involving large numbers. Therefore, whether it's a single round of multiplication in the single-cloud protocol or two rounds in the dual-cloud protocol, the processing times are relatively short.
3. **Cuckoo filter retrieval**: Although relatively quick, the retrieval operation for a Cuckoo filter involves memory access, which may make it slightly slower than simple arithmetic operations. The exact time required for this operation depends on the size of the filter and the efficiency of the implementation.
4. **Outputting y_j based on index j**: This operation involves retrieving an element from an array or list based on a specific index and is generally very fast, primarily limited by memory access speeds.

After a detailed analysis and comparison, this section has conducted a thorough exploration of the key computational tasks executed by the client across the three different protocols. These tasks include modular exponentiation, multiplication operations, Cuckoo filter retrieval, and data retrieval based on an index. By assessing these types of computations and their specific time consumptions, the following conclusions can be drawn:

1. **Unbalanced PSI protocol based on Cuckoo filter**: primarily relies on two rounds of modular exponentiation, which are computation-intensive, especially when dealing with large numbers.
2. **Unbalanced PSI protocol based on single-cloud assistance**: by executing a single round of multiplication and an index-based data retrieval process, it significantly alleviates the computational burden on the client.
3. **Unbalanced PSI protocol based on dual-cloud assistance**: Includes two rounds of multiplication operations and an index-based data retrieval process, also aiming to

distribute the computational pressure on the client. Although it involves two rounds of multiplication, due to the inherent efficiency of the operation, the total processing time remains within an acceptable range.

Through the meticulous assessment of each protocol's computational types and their time consumptions, it is evident that both the unbalanced PSI protocol based on single-cloud assistance and unbalanced PSI protocol based on dual-cloud assistance exhibit excellent performance in reducing the client's computational burden, particularly in the efficient execution of multiplication operations and data retrieval. In contrast, the unbalanced PSI protocol based on Cuckoo filter, while potentially offering stronger security provisions, shows some deficiencies in efficiency and timeliness. Therefore, when choosing an appropriate protocol, a balance should be struck based on actual performance requirements and security needs.

7.5.2. Protocol Running Time

After introducing all three protocols, this section primarily discusses the running times of the protocols. The running time of a protocol is an important benchmark for evaluation in this paper because it directly reflects the protocol's efficiency in practical operations. The factors affecting the running time of the protocol include computational time and communication time. As shown in Table 5, experiments were conducted to record the running times of the protocols under various data volume combinations. Table 5 also places the running times of the unbalanced PSI protocol based on Cuckoo filter, unbalanced PSI protocol based on single-cloud assistance, and unbalanced PSI protocol based on dual-cloud assistance side by side for comparative analysis. Here, Protocol 1 refers to the unbalanced PSI protocol based on Cuckoo filter, Protocol 2 refers to the unbalanced PSI protocol based on single-cloud assistance, and Protocol 3 refers to the unbalanced PSI protocol based on dual-cloud assistance.

Table 5. Running times of the three protocols under different data volume combinations.

Data Volume	Protocol 1 Running Time (s)	Protocol 2 Running Time (s)	Protocol 3 Running Time (s)
$2^{10}\|2^{15}$	0.1539	0.1543	0.1551
$2^{10}\|2^{17}$	0.1569	0.1573	0.1586
$2^{10}\|2^{20}$	0.1616	0.1611	0.1635
$2^{10}\|2^{25}$	0.1693	0.1683	0.1701
$2^{15}\|2^{15}$	4.9239	3.8223	4.3707
$2^{15}\|2^{17}$	5.0232	3.9145	4.4709
$2^{15}\|2^{20}$	5.1709	4.0267	4.6001
$2^{15}\|2^{25}$	5.4172	4.2233	4.8206
$2^{17}\|2^{17}$	20.0930	15.6768	17.8801
$2^{17}\|2^{20}$	20.6841	16.1281	18.4004
$2^{17}\|2^{25}$	21.6690	16.8939	19.2802
$2^{20}\|2^{20}$	165.4731	129.0516	147.2608
$2^{20}\|2^{25}$	173.3531	135.2534	154.3003

It is noteworthy that the preprocessing stages of all three protocols can be completed offline, meaning they do not directly contribute to online operation delays. Therefore, the recorded running times in this paper refer to the total time of all processes excluding preprocessing. Specifically, in Protocol 1, this primarily refers to the total duration of the intersection process; in Protocols 2 and 3, it refers to the total duration of both the outsourcing and intersection processes.

Through experimental analysis, the paper draws the following conclusions: At smaller data volumes, the performance differences between the three protocols are not significant. However, as the data volume increases, the differences in running times between the protocols become apparent. Generally, the unbalanced PSI protocol based on Cuckoo

filter tends to have the longest running time, while the unbalanced PSI protocol based on single-cloud assistance has the shortest running time, and the performance of the unbalanced PSI protocol based on dual-cloud assistance is in the middle. This phenomenon can be explained by the complexity of data handling and the differences in communication overhead among the protocols. The unbalanced PSI protocol based on Cuckoo filter, due to its direct and unoptimized calculations, is less efficient when handling large volumes of data. Nevertheless, at very small data volumes, where the proportion of communication time is relatively high, the impact of data transmission costs on total running time becomes significant. In such cases, the unbalanced PSI protocol based on Cuckoo filter does not necessarily appear inefficient because other protocols might be even less efficient in data transmission. Especially in environments with poor network conditions or limited data transfer rates, the lower communication demands of the unbalanced PSI protocol based on Cuckoo filter might, in some cases, lead to better performance.

Moreover, the running times of the unbalanced PSI protocol based on single-cloud assistance and the unbalanced PSI protocol based on dual-cloud assistance are significantly reduced through distributed computing and the use of cloud resources, especially when dealing with large-scale datasets. In summary, choosing the appropriate protocol requires a comprehensive consideration of factors such as data volume, computational resources, and network environment. In practical applications, understanding the performance characteristics and suitable scenarios of each protocol is crucial for optimizing data processing workflows and enhancing efficiency.

7.6. Summary of This Chapter

The protocol leverages the computational and storage resources of two cloud servers, significantly reducing the burden on the client by lowering its computational and storage requirements and enhancing the system's efficiency and availability. Through distributed computing and security measures such as homomorphic encryption, it ensures the safety and privacy of data during transmission and processing, adequately protecting sensitive information of both the client and the database server. This solution not only improves the operational efficiency of devices with limited resources but also effectively prevents collusion attacks. Consequently, the unbalanced PSI protocol based on dual-cloud assistance excels in PSI operations, particularly in applications like private contact discovery, demonstrating both high efficiency and security.

7.7. Extensions

To enhance the practicality of the scheme, this section will explore two key aspects from an engineering practice perspective: the design of the PSI network and the design of the data update mechanism.

First, the design of the PSI network focuses on building an efficient, secure, and scalable network architecture to support large-scale PSI computations.

Second, the design of the data update mechanism involves a method to update the datasets stored on the cloud servers without interrupting the service. This is particularly crucial for PSI computation scenarios that require frequent data updates.

7.7.1. PSI Network

As previously described, the client delegates PSI computations to two cloud servers. In practice, a vast network of cloud servers can be built to support this delegation. The basic system description is as follows.

- Access and authentication of cloud servers: Any server can apply to become a cloud server, also known as a server assistant. These servers must undergo a series of certification processes (including hardware performance verification, security vulnerability scanning, and compliance checks) to ensure they meet security and performance standards. Servers that pass the certification but later violate regulations will be blacklisted and removed. The system maintains platform security and trust through mechanisms

such as regular security scans and real-time monitoring, with any violations leading to the immediate removal and further investigation of the server.

- Mechanism for selecting Server assistants: When needing to perform PSI, clients choose two cloud servers based on their performance (such as processing power, storage capacity, and network bandwidth), stability, security capabilities, and compliance with regulations, among other hard and soft factors. Cloud servers with high availability promises are preferred to minimize the risk of failures.
- Execution mechanism for PSI operations: The PSI network supports client flexibility and system scalability; clients can execute PSI on different database servers by merely changing $E_{pk}(a)$, without needing to redesign the entire system. This design enhances client flexibility and the system's efficiency, reliability, and security.

7.7.2. Summary of the PSI Network

This system design not only achieves the delegation of PSI computations but also introduces multiple cloud servers into the network, thereby enhancing the system's flexibility and stability. Additionally, by implementing authentication and maintaining a blacklist for cloud servers, the system can better guarantee the credibility of the cloud servers, enhancing overall security. This flexible yet secure system design provides clients with more options and makes PSI operations more adaptable to various practical requirements.

In summary, under the existing framework, clients can delegate computing and storage tasks to different cloud servers and perform PSI operations on various database servers by using different random numbers and different $E_{pk}(a)$. This method allows clients to more flexibly use multiple resource nodes and optimize task distribution, thereby further enhancing the overall performance and security of the privacy protection scheme.

7.7.3. Data Updates

To further enhance the practicality of the scheme, this paper also designs a data update mode compatible with the scheme, making the overall scheme more practical and reliable.

- **Data Updates on the Database Server Side:**

  As shown in Figure 6, the update details of the database server are as follows:
  Definition of main participants and related symbols:

  1. database server: represents the database server that wants to encrypt and upload updated data to cloud server H_2.
  2. cloud server H_2): represents the cloud-assisted server H_2 that assists the database server in completing update operations.
  3. Z represents the set of data to be updated, z_k represents the k-th element of Z.
  4. ω represents the load factor of the filter.
  5. z'_k represents the data z_k after encryption processing.
  6. update$_k$ represents the operation index, used to determine whether the update operation is an insertion or deletion.
  7. U represents the set of data sent by the database server to the cloud-assisted server H_2, u_k represents the k-th element of U.

  Update process:

  1. The database server has a set of elements Z it wants to insert or delete. These elements are blinded before being sent to cloud server H_2. Specifically, $z'_k = E_{pk_c}(a)^{\alpha * H(z_k)}$.
  2. In addition to sending the blinded elements, the database server also sends an identifier variable update$_k$ to inform the client whether the operation is an insertion or a deletion.
  3. During an insertion operation, H_2 first checks whether the current filter's load factor ω exceeds 0.95.

4. If the load factor ω is greater than 0.95, then H_2 must request the database server to generate a new filter using all elements to maintain high spatial and lookup efficiency of the filter.
5. If the load factor ω is less than or equal to 0.95, then H_2 can directly insert the element into the current filter CF.
6. In a deletion operation, H_2 removes the specified element from the filter CF, a process that does not require generating a new filter.

This section introduces the data update process for the database server under the unbalanced PSI protocol based on single-cloud assistance. This series of update operations allows the database server to flexibly handle the insertion and deletion of elements based on the current state of the filter, ensuring the system's efficiency and accuracy.

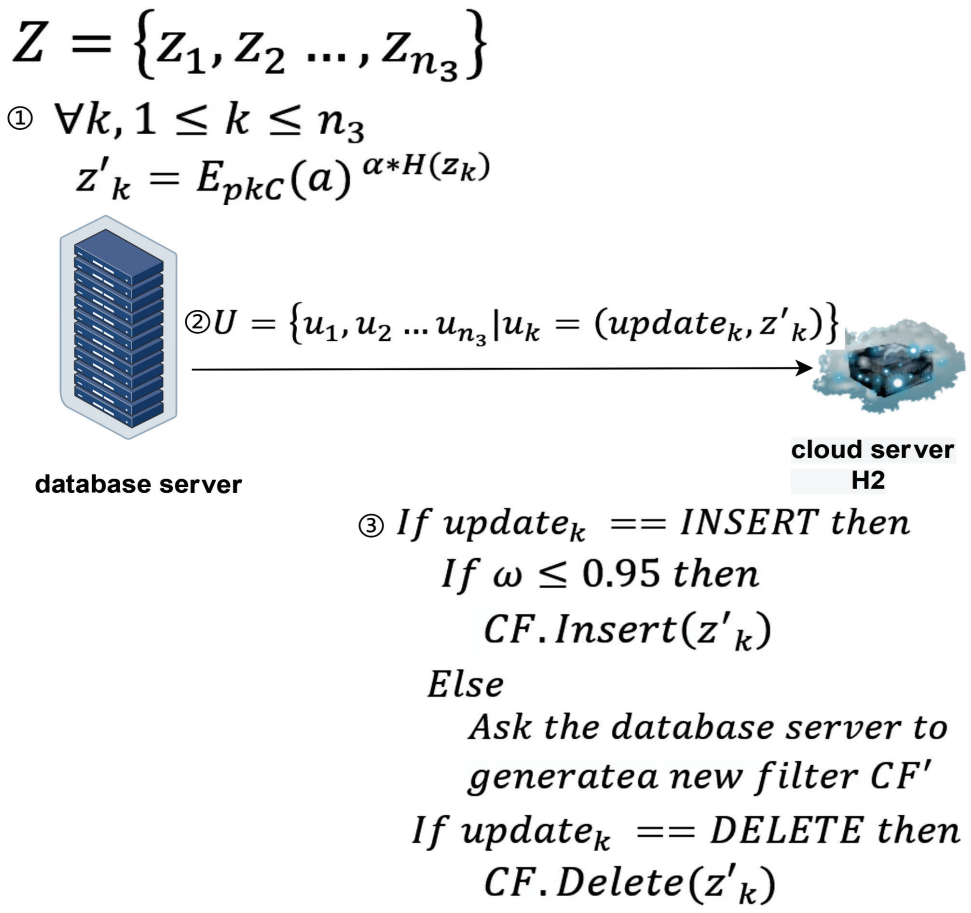

Figure 6. Data Updates on the Database Server Side.

- **Data updates on the client side:** As shown in Figure 7, the update details of the database server are as follows:
  Definition of main participants and related symbols:
  1. client: represents the client who wants to perform data updates.
  2. cloud server H_1: represents the cloud-assisted server H_1 that assists the database server in completing update operations.
  3. cloud server H_2 represents the cloud-assisted server H_2 that assists the database server in completing update operations.
  4. Z represents the set of data to be updated, z_k represents the k-th element of Z.
  5. z'_k represents the data after being processed by the hash function H.
  6. k represents the data index, used to determine the type of update, either insertion or deletion, and to retrieve the updated data based on the index.

7. When adding data, date_k represents the data processed through the dual-cloud scheme and sent to the two cloud-assisted servers. When deleting, date_k is null.
8. V represents the set of data sent by the database server to the cloud-assisted server H_1, v_k represents the k-th element of V.
9. V' represents the set of data sent by the database server to the cloud-assisted server H_2, v'_k represents the k-th element of V'.

Update process:

1. The client has a set of elements Z it wants to insert or delete. In both cases, the client blinds each element and sends them to H_1 and H_2 respectively.
2. The client sends a data index K to inform the cloud servers about the type of update, whether it is an insertion or a deletion. If the index is less than n_2, it indicates a deletion operation. In this case, data_k is null, and H_1 and H_2 delete the corresponding data based on the index.
3. If the index is greater than n_2, it indicates an addition operation, and the corresponding calculation results and index are saved.
4. After completing a batch of deletion and addition operations, the relative order of the indices also needs to be adjusted. The update process is illustrated in Figure 5.

This section introduces a client data update process based on dual-cloud assistance, designed to enhance the database's dynamic management capabilities while ensuring data privacy and efficiency. This update protocol supports both data insertion and deletion operations, and through the cooperation of cloud-assisted servers H_1 and H_2, it optimizes the speed and security of client data updates.

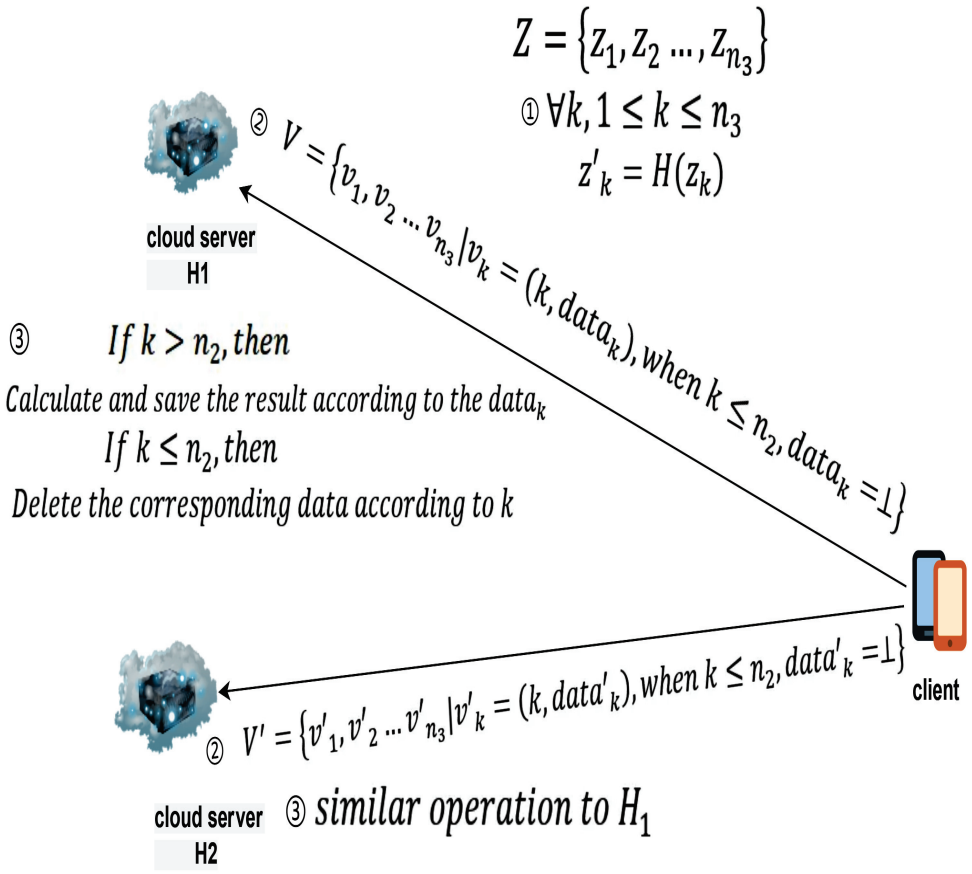

Figure 7. Data updates on the client's side.

- **Summary of data updates:** This section has explored two key data update processes based on the unbalanced PSI protocol based on dual-cloud assistance the database server update process and the client update process. Both processes are designed to efficiently handle data insertions and deletions while ensuring data security, and to use cloud server resources to optimize overall operation efficiency.

8. Conclusions and Future Work

8.1. Work Summary

Privacy computing is a technology framework aimed at protecting individual privacy during the process of data use and sharing. It ensures that data can still be effectively utilized without disclosing specific content through various algorithms and protocols. Among many applications of privacy computing, Privacy Set Intersection (PSI) is a common requirement, which allows two or more parties to compute the intersection of their datasets without revealing their exclusive data to each other.

Traditional PSI protocols are primarily designed for cases where datasets are relatively balanced in size, which often does not apply in real scenarios. In many practical situations, the size disparity between participants' datasets is significant, necessitating the use of unbalanced PSI protocols. These protocols are specifically designed to handle such disparities, optimizing computational efficiency and privacy protection to cater to a wider range of practical needs. By adopting unbalanced PSI protocols, not only is the processing efficiency improved, but more precise control over data protection is also offered, thus finding broader application in various data-sensitive industries. This paper proposes three protocols: the unbalanced PSI protocol based on Cuckoo filter, the unbalanced PSI protocol based on single-cloud assistance, and the unbalanced PSI protocol based on dual-cloud assistance. Here, the unbalanced PSI protocol based on Cuckoo filter addresses performance issues of traditional PSI protocols in handling unbalanced datasets. On this basis, the unbalanced PSI protocol based on single-cloud assistance transfers most of the computational and storage burdens from the client to the cloud, enhancing practicality. Faced with the possibility of collusion attacks, the unbalanced PSI protocol based on dual-cloud assistance employs security mechanisms such as homomorphic encryption to effectively resist these attacks. The main contributions of this paper are summarized as follows:

1. Addressing the shortcomings of traditional PSI protocols when dealing with significant data size disparities among participants, this paper proposes the first protocol, namely the unbalanced PSI protocol based on Cuckoo filter. This protocol successfully constructs a novel PSI approach through encrypted exchanges and using Cuckoo filter for private information retrieval.
2. Given the complexities of cryptographic operations and storage demands in the unbalanced PSI protocol based on Cuckoo filter, this paper introduces a unbalanced PSI protocol based on single-cloud assistance. This protocol successfully offloads most of the client's computational and storage burdens onto the cloud.
3. In response to potential collusion between the cloud and database servers in the unbalanced PSI protocol based on single-cloud assistance, this paper proposes an unbalanced PSI protocol based on dual-cloud assistance with security mechanisms like homomorphic encryption, which effectively prevents collusion attacks while offloading computational and storage burdens.
4. Concerning practical issues in the unbalanced PSI protocol based on dual-cloud assistance, this paper also introduces a conceptually meaningful PSI network and a data update mode tailored for the unbalanced PSI protocol based on dual-cloud assistance.

8.2. Protocol Summary

As shown in Table 6, this section provides a comprehensive summary and recommendations for the three protocols discussed in this paper. The unbalanced PSI protocol based on Cuckoo filter offers high security but involves significant computational and storage demands, making it suitable for clients with strong computational and storage resources.

The unbalanced PSI protocol based on single-cloud assistance, while being the fastest and offloading computational burdens to the cloud, poses security risks as it cannot withstand collusion attacks, making it appropriate for scenarios where the cloud is fully trusted. The unbalanced PSI protocol based on dual-cloud assistance offers an ideal balance of runtime, security, and efficiency, making it the most versatile and practical option.

Table 6. Overall performance summary of the three protocols.

Protocol	Security	Client Storage and Computational Burden	Runtime
Unbalanced PSI protocol based on Cuckoo filter	High security (no collusion attacks)	Requires storing Cuckoo filter and intensive computation	Longest
Unbalanced PSI protocol based on single-cloud assistance	Security risks (cannot resist collusion attacks)	Shifted to cloud server	Fastest
Unbalanced PSI protocol based on dual-cloud assistance	High security (can resist collusion attacks)	Shifted to cloud server	Moderate

8.3. Future Outlook

Although the protocols proposed in this document are applicable in most scenarios, there are still several aspects that could be optimized for future development:

1. All protocols are designed for two-party unbalanced PSI. Extending these protocols to multi-party scenarios is an important future direction, given the practical needs for multi-party computations.
2. The protocols are developed under a semi-honest security model. Extending their robustness to malicious models, where adversaries may actively attempt to undermine the protocols, represents a crucial area for further research.
3. The current protocols focus solely on set intersection. In practical applications, there may be requirements to perform further computations on the intersection results. Developing functionalities to support such computations post-intersection is another significant direction for future work.

Author Contributions: Conceptualization, S.D.; Methodology, J.W.; Software, Q.F.; Validation, J.W.; Formal analysis, Q.F.; Investigation, Q.F.; Resources, W.T. and J.W.; Data curation, S.D.; Writing—original draft, Q.F. and S.D.; Writing—review & editing, Q.F.; Visualization, S.D.; Supervision, W.T. and S.D.; Project administration, Q.F. and W.T.; Funding acquisition, W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guangdong Key Laboratory of Data Security and Privacy Preserving: 2023B1212060036.

Data Availability Statement: The original data for this article were generated and recorded by ourselves.The data for this article does not originate from a publicly available repository.The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bald, P.; Baronio, R.; Cristofaro, E.; Gasti, P.; Tsudik, G. Efficient and secure testing of fully-sequenced human genomes. *Biol. Sci. Initiat.* **2000**, *470*, 7–10.
2. Chen, H.; Laine, K.; Rindal, P. Fast private set intersection from homomorphic encryption. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–3 November 2017; pp. 1243–1255.
3. Nagaraja, S.; Mittal, P.; Hong, C.Y.; Caesar, M.; Borisov, N. BotGrep: Finding P2P Bots with Structured Graph Analysis. In Proceedings of the 19th USENIX Security Symposium (USENIX Security 10), Washington, DC, USA, 11–13 August 2010.
4. Chen, H.; Huang, Z.; Laine, K.; Rindal, P. Labeled PSI from fully homomorphic encryption with malicious security. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018; pp. 1223–1237.

5. Kamara, S.; Mohassel, P.; Raykova, M.; Sadeghian, S. Scaling private set intersection to billion-element sets. In Proceedings of the Financial Cryptography and Data Security: 18th International Conference, FC 2014, Christ Church, Barbados, 3–7 March 2014; Revised Selected Papers; Springer: Berlin/Heidelberg, Germany, 2014; pp. 195–215.
6. Pinkas, B.; Schneider, T.; Weinert, C.; Wieder, U. Efficient circuit-based PSI via cuckoo hashing. In Proceedings of the 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, 19–23 May 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 125–157.
7. Li, W.; Liu, J.; Zhang, L.; Wang, Q.; He, C. A Survey on Set Intersection Computation for Privacy Protection. *J. Comput. Res. Dev.* **2022**, *59*, 1782–1799.
8. Meadows, C. A More Efficient Cryptographic Matchmaking Protocol for Use in the Absence of a Continuously Available Third Party. In Proceedings of the 7th IEEE Symposium on Security and Privacy, Los Alamitos, CA, USA, 7–9 April 1986; p. 134.
9. Huberman, B.; Franklin, M.; Hogg, T. Enhancing Privacy and Trust in Electronic Communities. In Proceedings of the 1st ACM Conference on Electronic Commerce, Denver, CO, USA, 3–5 November 1999; pp. 78–86.
10. DeCristofaro, E.; Tsudik, G. Experimenting with Fast Private Set Intersection. In *International Conference on Trust and Trustworthy Computing*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 55–73.
11. Pinkas, B.; Schneider, T.; Zohner, M. Faster Private Set Intersection Based on OT Extension. In Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, 20–22 August 2014; pp. 797–812.
12. Freedman, M.; Nissim, K.; Pinkas, B. Efficient Private Matching and Set Intersection. In Proceedings of the 23rd International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, 2–6 May 2004; Springer: Berlin/Heidelberg, Germany, 2004.
13. Freedman, M.J.; Hazay, C.; Nissim, K.; Pinkas, B. Efficient Set Intersection with Simulation-Based Security. *J. Cryptol.* **2016**, *29*, 115–155. [CrossRef]
14. Abadi, A.; Terzis, S.; Dong, C. O-PSI: Delegated Private Set Intersection on Outsourced Datasets. In Proceedings of the 30th IFIP International Information Security and Privacy Conference, Hamburg, Germany, 26–28 May 2015; pp. 3–17.
15. Kissner, L.; Song, D. Privacy-Preserving Set Operations. In Proceedings of the 25th Annual International Cryptology Conference, Santa Barbara, CA, USA, 14–18 August 2005; pp. 241–257.
16. Jarecki, S.; Liu, X. Efficient Oblivious Pseudorandom Function with Applications to Adaptive OT and Secure Computation of Set Intersection. In Proceedings of the LNCS 5444: 6th Theory of Cryptography Conference, San Francisco, CA, USA, 15–17 March 2009; pp. 577–594.
17. Hazay, C.; Venkitasubramaniam, M. Scalable Multi-party Private Set-Intersection. In Proceedings of the 20th IACR International Workshop on Public Key Cryptography, Amsterdam, The Netherlands, 28–31 March 2017; pp. 175–203.
18. Dou, J.; Liu, X.; Wang, W. Efficient and Secure Calculation of Two-Party Sets in the Field of Rational Numbers. *Chin. J. Comput.* **2020**, *43*, 1397–1413.
19. Damgård, I.; Pastro, V.; Smart, N.; Zakarias, S. Multiparty Computation from Somewhat Homomorphic Encryption. In Proceedings of the 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, 19–23 August 2012; Lecture Notes in Computer Science; Safavi-Naini, R., Canetti, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 643–662.
20. Yao, A.C. Protocols for Secure Computations. In Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982), Chicago, IL, USA, 3–5 November 1982; pp. 160–164.
21. Micali, S.; Goldreich, O.; Wigderson, A. How to Play Any Mental Game. In Proceedings of the 19th ACM Symposium on Theory of Computing, New York, NY, USA, 25–27 May 1987; pp. 218–229.
22. Pinkas, B.; Schneider, T.; Segev, G.; Zohner, M. Phasing: Privateset intersectionusing permutation-basedhashing. In Proceedings of the 24th USENIX Security Symposium, Washington, DC, USA, 12–14 August 2015; USENIX Association: Berkeley, CA, USA, 2015; pp. 515–530.
23. Pinkas, B.; Schneider, T.; Tkachenko, O.; Yanai, A. Efficient circuit-based PSI with linear communication. In Proceedings of the 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, 10–14 May 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 122–153.
24. Huang, Y.; Evans, D.; Katz, J. Private Set Intersection: Are Garbled Circuits Better Than Custom Protocols? In Proceedings of the 19th Network and Distributed System Security Symposium, San Diego, CA, USA, 5–8 February 2012.
25. Naor, M.; Pinkas, B. Efficient oblivious transfer protocols. In Proceedings of the SODA, Washington, DC, USA, 7–9 January 2001; Volume 1, pp. 448–457.
26. Dong, C.; Chen, L.; Wen, Z. When private set intersection meets big data: An efficient and scalable protocol. In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, Berlin, Germany, 4–8 November 2013; pp. 789–800.
27. Rindal, P.; Rosulek, M. Improved private set intersection against malicious adversaries. In Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques, Paris, France, 30 April–4 May 2017; Springer International Publishing: Cham, Switzerland, 2017; pp. 235–259.
28. Zhang, E.; Liu, F.H.; Lai, Q.; Jin, G.; Li, Y. Efficient multi-party private set intersection against malicious adversaries. In Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop, London, UK, 11 November 2019; pp. 93–104.

29. Pinkas, B.; Rosulek, M.; Trieu, N.; Yanai, A. PSI from PaXoS: Fast, malicious private set intersection. In Proceedings of the 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, 10–14 May 2020; Springer: Cham, Switzerland, 2020; pp. 739–767.
30. Orrù, M.; Orsini, E.; Scholl, P. Actively secure 1-out-of-n OT extension with application to private set intersection. In Proceedings of the Cryptographers' Track at the RSA Conference, San Francisco, CA, USA, 14–17 February 2017; Springer: Cham, Switzerland, 2017; pp. 381–396.
31. Rindal, P.; Schoppmann, P. VOLE-PSI: Fast OPRF and Circuit-PSI from Vector-OLE. IACR Cryptology ePrint Archive. 2021. Available online: https://eprint.iacr.org/2021/266 (accessed on 22 April 2024).
32. Schoppmann, P.; Gascón, A.; Reichert, L.; Raykova, M. Distributed vector-OLE: Improved constructions and implementation. In Proceedings of the 26th ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 1055–1072.
33. Weng, C.; Yang, K.; Katz, J.; Wang, X. Wolverine: Fast, Scalable, and Communication-Efficient Zero-Knowledge Proofs for Boolean and Arithmetic Circuits. Cryptology ePrint Archive. 2020. Available online: https://eprint.iacr.org/2020/925 (accessed on 22 April 2024).
34. Wang, Z.; Ma, X. Blockchain-Based Unbalanced PSI with Public Verification and Financial Security. *Mathematics* **2024**, *12*, 1544. [CrossRef]
35. Ning, J.; Tan, Z.; Zhang, K.; Ye, W. Low Communication-Cost PSI Protocol for Unbalanced Two-Party Private Sets. *IET Inf. Secur.* **2024**, *2024*, 6052651. [CrossRef]
36. Zhao, Q.; Jiang, B.; Zhang, Y.; Wang, H.; Mao, Y.; Zhong, S. Unbalanced private set intersection with linear communication complexity. *Sci. China Inf. Sci.* **2024**, *67*, 1–15. [CrossRef]
37. Tan, W.; Du, S.; Weng, J. Efficient Cryptographic Solutions for Unbalanced Private Set Intersection in Mobile Communication. *Preprints* **2024**, 2024041701. [CrossRef]
38. Yang, X.; Cai, L.; Wang, Y.; Sun, L.; Hu, J. Efficient Unbalanced Quorum PSI from Homomorphic Encryption. *Cryptol. Eprint Arch.* 2024, *preprint*.
39. Chen, Y.; Wu, A.; Yang, Y.; Xin, X.; Song, C. Efficient Verifiable Cloud-Assisted PSI Cardinality for Privacy-Preserving Contact Tracing. *IEEE Trans. Cloud Comput.* **2024**, *12*, 251–263. [CrossRef]
40. Van Baarsen, A.; Stevens, M. Amortizing Circuit-PSI in the Multiple Sender/Receiver Setting. *Cryptol. Eprint Arch.* 2024, *preprint*.
41. Bienstock, A.; Patel, S.; Seo, J.Y.; Yeo, K. Batch PIR and Labeled PSI with Oblivious Ciphertext Compression. *Cryptol. Eprint Arch.* 2024, *preprint*.
42. Hao, M.; Liu, W.; Peng, L.; Li, H.; Zhang, C.; Chen, H.; Zhang, T. Unbalanced Circuit-PSI from Oblivious Key-Value Retrieval. *Cryptol. Eprint Arch.* 2023, *preprint*.
43. Son, Y.; Jeong, J. PSI with computation or Circuit-PSI for Unbalanced Sets from Homomorphic Encryption. In Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, Melbourne, Australia, 10–14 July 2023; pp. 342–356.
44. Yang, Y.; Dong, X.; Shen, J.; Cao, Z.; Yang, Y.; Zhou, J.; Fang, L.; Liu, Z.; Ge, C.; Su, C.; et al. MDPPC: Efficient Scalable Multiparty Delegated PSI and PSI Cardinality. In Proceedings of the 2023 20th Annual International Conference on Privacy, Security and Trust (PST), Copenhagen, Denmark, 21–23 August 2023; pp. 1–7.
45. Papafragkos, P.; Gavalas, I.; Raptopoulos, I.; Chasalevris, A. Optimizing energy dissipation in gas foil bearings to eliminate bifurcations of limit cycles in unbalanced rotor systems. *Nonlinear Dyn.* **2023**, *111*, 67–95. [CrossRef]
46. Arivumani, V.; Balaraman, S. Angular symmetrical components-based anti-islanding method for solar photovoltaic-integrated microgrid. *Automatika* **2023**, *64*, 1–21. [CrossRef]
47. Berenjian, S. Encryption-Based Secure Protocol Design for Networks. Ph.D. Thesis, Stevens Institute of Technology, Hoboken, NJ, USA, 2023.
48. Hill, K. Facebook Figured Out My Family Secrets, And It Won't Tell Me How. *Gizmodo* **2017**. Published on 25 August 2017. Available online: https://gizmodo.com/facebook-figured-out-my-family-secrets-and-it-wont-tel-1797696163 (accessed on 22 April 2024).
49. Marlinspike, M. Private Contact Discovery for Signal. 2017. Available online: https://signal.org/blog/private-contact-discovery (accessed on 26 September 2017).
50. Mittal, P.; Papamanthou, C.; Song, D. Preserving Link Privacy in Social Network Based Systems. In Proceedings of the NDSS, San Diego, CA, USA, 24–27 February 2013.
51. Abebe, R.; Nakos, V. *Private Link Prediction in Social Networks*; Technical Report; Harvard University: Cambridge, MA, USA, 2014.
52. Karwa, V.; Raskhodnikova, S.; Smith, A.; Yaroslavtsev, G. Private Analysis of Graph Structure. *Proc. VLDB Endow.* **2011**, *4*, 1146–1157. [CrossRef]
53. Dwork, C. A Firm Foundation for Private Data Analysis. *Commun. ACM* **2011**, *54*, 86–95. [CrossRef]
54. Erlingsson, Ú.; Pihur, V.; Korolova, A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), Scottsdale, AZ, USA, 3–7 November 2014.
55. Brendel, W.; Han, F.; Marujo, L.; Jie, L.; Korolova, A. Practical privacy-preserving friend recommendations on social networks. In Proceedings of the Companion Proceedings of the The Web Conference 2018, Gothenburg, Sweden, 27 May–3 June 2018; pp. 111–112.

56. Su, G.; Xu, M. A Survey on Secure Multi-party Computation Technology and Applications. *Inf. Commun. Technol. Policy* **2019**, *5*, 19–22.
57. Wang, H.; Dai, H.; Chen, S.; Chen, Z.; Chen, G. A Survey of Filter Data Structures. *Comput. Sci.* **2024**, *51*, 35–40.
58. Yu, M.; Fabrikant, A.; Rexford, J. BUFFALO: Bloom filter forwarding architecture for large organizations. In Proceedings of the International Conference on Emerging Networking Experiments and Technologies, Rome, Italy, 1–4 December 2009; pp. 313–324.
59. Li, P.; Luo, B.; Zhu, W.; Xu, H. Cluster-based distributed dynamic cuckoo filter system for Redis. *Int. J. Parallel Emergent Distrib. Syst.* **2020**, *35*, 340–353. [CrossRef]
60. Wang, F.; Chen, H.; Liao, L.; Zhang, F.; Jin, H. The power of better choice: Reducing relocations in cuckoo filter. In Proceedings of the International Conference on Distributed Computing Systems, Dallas, TX, USA, 7–9 July 2019; pp. 358–367.
61. Gu, R.; Li, S.; Dai, H.; Wang, H.; Luo, Y.; Fan, B.; Basat, R.B.; Wang, K.; Song, Z.; Chen, S.; et al. Adaptive online cache capacity optimization via lightweight working set size estimation at scale. In Proceedings of the USENIX Annual Technical Conference, Boston, MA, USA, 10–12 July 2023; pp. 467–484.
62. Reviriego, P.; Martínez, J.; Larrabeiti, D.; Pontarelli, S. Cuckoo Filters and Bloom Filters: Comparison and Application to Packet Classification. *IEEE Trans. Netw. Serv. Manag.* **2020**, *17*, 2690–2701. [CrossRef]

information

MDPI

Review

Digital Transformation in Energy Sector: Cybersecurity Challenges and Implications

Saqib Saeed [1,*], Hina Gull [1], Muneera Mohammad Aldossary [1,*], Amal Furaih Altamimi [1], Mashael Saeed Alshahrani [1], Madeeha Saqib [1], Sardar Zafar Iqbal [1] and Abdullah M. Almuhaideb [2]

[1] Saudi Aramco Cybersecurity Chair, Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia; hgull@iau.edu.sa (H.G.); 2240500117@iau.edu.sa (A.F.A.); 2240500111@iau.edu.sa (M.S.A.); mssaeed@iau.edu.sa (M.S.); saiqbal@iau.edu.sa (S.Z.I.)

[2] Saudi Aramco Cybersecurity Chair, Department of Networks and Communications, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia; amalmuhaideb@iau.edu.sa

* Correspondence: sbsaed@iau.edu.sa (S.S.); 2240500109@iau.edu.sa (M.M.A.)

Abstract: Digital transformation in energy sector organizations has huge benefits but also exposes them to cybersecurity challenges. In this paper, we carried out a systematic literature review on cybersecurity challenges and issues in the energy domain. Energy-associated assets are very critical for any nation and cyber-attacks on these critical infrastructures can result in strategic, financial, and human losses. We investigated research papers published between 2019 and 2024 and categorized our work into three domains: oil and gas sector, the electricity sector, and the nuclear energy sector. Our study highlights that there is a need for more research in this important area to improve the security of critical infrastructures in the energy sector. We have outlined research directions for the scientific community to further strengthen the body of knowledge. This work is important for researchers to identify key areas to explore as well as for policymakers in energy sector organizations to improve their security operations by understanding the associated implications of cybersecurity.

Keywords: cybersecurity; digital transformation; energy sector automation; risk assessment; operations

1. Introduction

The energy sector is not only an important sector of any economy but is also an essential component of modern human life. It is difficult to imagine life without electricity and fuel as modern-day work infrastructures are heavily dependent on energy. According to the International Energy Agency (IEA), global demand has seen a growth of 2.2% in 2023; however, a 3.4% annual growth is expected by 2026. Although there is a slight decrease in electricity consumption in the US and EU, emerging economies like China and India have seen a growth in electricity consumption. Technological advancements such as artificial intelligence, data centers, and cryptocurrency are estimated to double their energy consumption by 2026 [1]. These statistics highlight the need for more energy production sources and efficient management of energy production, distribution, and use.

Digital transformation has advocated for fostering digital technologies to improve business processes and customer satisfaction and many sectors have benefited from digital transformation initiatives. Similarly, digital transformation in the energy sector has been explored and it has been highlighted that digitalization can help reduce energy consumption and optimize energy sector operations [2–4]. Nazari and Musilek [5] also highlighted that digital transformation in the energy sector results in cost reduction, efficiency, and enhanced customer experience. In another study, Oudina et al. [6] highlighted that petroleum and natural gas are important forms of energy that support the expansion of numerous

other businesses, as well as many aspects of contemporary living and the world economy. The ecological impact of the oil and gas industry is changing due to the Petroleum Cyber-Physical System (CPS). Petroleum CPS efficiency approaches aid in an international assessment of the oil field by taking into account the amount of information on output generated by a drilling site. The energy sector is exposed to a number of dangers that have the potential to damage the natural world, interrupt vital power lines, and trigger an economic calamity. These dangers include human errors, environmental hazards, cyberattacks, and disruptions in connectivity. The focus of the scientific community is on the development of a self-aware and contemporary CPS. The research on petroleum and natural gas lacks the definition of threats, the reasons behind reservations of all kinds, and a workable defense strategy. Moreover, a thorough investigation of cyber security for oil and gas industries is still lacking. In a reported study, authors discussed the basic trust issues with CPS, along with how they apply to the oil and gas sector. They divided trust-related issues into functional, human, business, and trust categories. The issues were outlined as a group of characteristics and showed how they are related. This study found that recognizing and resolving issues in the oil and gas sector is a critical first step in implementing risk prevention, protection, and mitigation strategies, and is a vital tool for enhancing CPS reliability and quality in this important economic domain. In order to measure the readiness level for digital transformation in the energy sector, different assessment methodologies have been developed that can help in measuring the readiness of the energy sector in different countries [7]. Such assessment frameworks can help in prioritizing the need of improvement in the energy sector infrastructure for policymakers and higher management.

COVID-19 has accelerated the digital transformation rate in industries to achieve business continuity by adopting modern technologies such as big data analytics, cloud computing, internet of things, artificial intelligence, etc. However, associated cybersecurity implications of digital transformation have also increased, posing a security threat for organizations [8–10]. Recent cybersecurity attacks such as Stuxnet [11], colonial pipeline hack [12], Dragonfly attacks [13], cyberattacks on Saudi Aramco [14], and cyberattacks on smart meters in Puerto Rico [15] further strengthen the importance of cybersecurity in energy sector [16].

In this paper, we have conducted a systematic literature review, which mainly focuses on cybersecurity implications in oil and gas and the electricity and nuclear energy sectors. This paper presents recommendations for further research in this domain. The findings will help in formulating appropriate security policies to enable the benefits of digital transformation in the energy sector. This paper is structured as follows: Section 2 explains the procedures for selecting primary studies for systematic analysis and Section 3 explains the summaries of selected papers. Section 4 highlights the findings of the study and Section 5 concludes the analysis and makes some recommendations for further studies.

2. Materials and Methods

In this section, we explain the methodology adopted to extract the research papers for our study. We performed a systematic literature review using the PRISMA guidelines [17], we used the Google Scholar [18] database as our main repository for scientific papers. We queried the repository by different keywords which were chosen to facilitate the extraction of research articles related to our topic. The search terms used were (digital transformation in energy sector) AND (cyber security), (digital transformation in oil and gas) AND (cybersecurity), (Cybersecurity in Electricity), (Digital Transformation in Nuclear energy) AND (cyber security), (energy sector) AND (cybersecurity). For each search result, first, we applied the duration filter from 2019 to 2024 and then we shortlisted the first 30 hits for each search results, resulting in a total of 150 papers. Based on this we carried out title filtering, then we filtered based on the abstract, and finally we performed filtering based on the contents of the paper. Our qualification criteria for research papers included the following:

- The paper is published in the time period of 2019–2024;
- The paper is focusing on both digital transformation and cybersecurity implications in energy sector;
- The paper is available in the English language;
- The paper is a not a review paper, book, or thesis.

As shown in Figure 1, after the final content filtering, we were left with 26 papers which are included in this study. The distribution of papers based on the publication year is shown in Figure 2, where we had 1 paper from 2019, 4 papers published in 2020, 3 papers published in 2021 and 2022 each, 11 papers from 2023, and 4 papers were published in 2024.

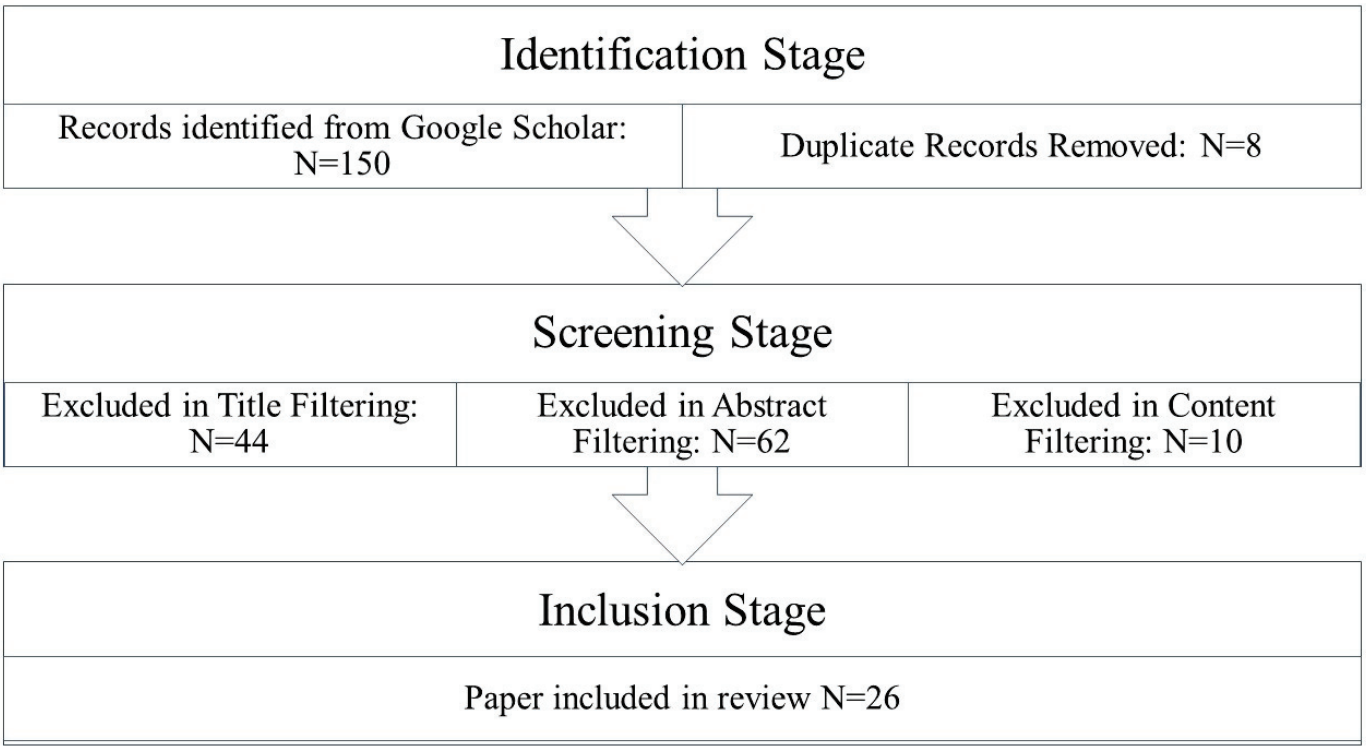

Figure 1. Research Methodology for Our Review Paper.

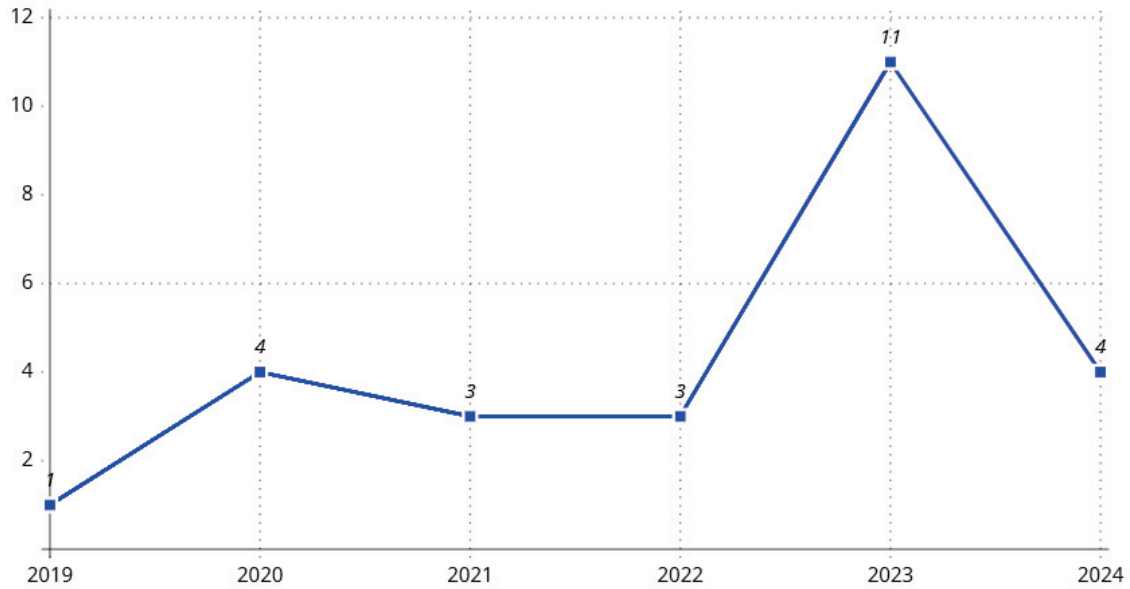

Figure 2. Distribution of Publication Years of Selected Papers in Review.

3. Results

In this section, we present the findings of our study grounded in the literature focusing on cybersecurity implications in Energy sector. There is a sparse body of knowledge focusing on recommendations to secure the industry [19].

3.1. Cybersecurity Implications in Oil and Gas Industry

Nowadays, cyberattacks directed at oil and gas firms have increased in onshore and offshore installations. Many cybercriminals and hackers target their cyberattacks on the oil and gas sector. Nowadays, the oil and gas sector has become more and more advanced in terms of technology. Activities are being digitalized at an increasing rate; sensors are being used. Although this increases productivity, it also increases the susceptibility of networks to cyberattacks [20]. Since offshore oil production typically takes place in isolated areas, remote access and control are necessary. Industrial Cyber-Physical Systems (ICPS), Supervisory, Control, and Data Acquisition (SCADA) systems, and Industrial Internet of Things (IIoT) technologies are integrated to achieve this system. The ecosystem, marine life, and worker safety could all be severely impacted by a successful cyberattack against an offshore oil and gas asset [21]. Frederick et al. [22] highlighted the investigation of cybersecurity scrutinization and management opportunities in the IIOT systems. Cybersecurity monitoring and controls are essential because of the continuous cyberattacks on these integrated IIoT systems. They outlined five strategies that support cybersecurity monitoring and control in the context of the industrialized Internet of Things and adjacent sectors. They looked at past and present incidents involving cyber threats and cyberattacks in industrial IoT systems to acquire data for their study. They also discussed novel ideas, applications, best practices, and systems for monitoring and control that, when put into effect, will benefit other sectors of the economy.

Villarreal et al. [23] have emphasized that international cyberattacks increased by 50% in 2021 as a result of COVID-19. The conflict in Ukraine has also made matters worse since 2022, especially for the oil and gas sector. They described that the industry must take a proactive approach to cybersecurity concerns by integrating behavioral and technology safeguards. To safeguard a natural gas company's technical structure, they suggested a model for cybersecurity based on the NIST CSF (National Institute of Standards and Technology Cybersecurity Framework) and CIS CSC (Center for Internet Security Critical Security Controls). The four steps of this paradigm are environmental analysis, scope and risk appetite, control design, and development. Through the evaluation of the applied techniques and employee surveys, this framework proved to be effective, producing a 92.69% efficacy and an 81.55% acceptance rate, meeting level two of the NIST CSF, and producing the desired result.

Houmb et al. [24] argued that higher levels of system integration and connectivity are necessary for intelligent automated industrial process control. An increasing danger of cyberattacks for Industrial Control Systems (ICS) and other Operational Technology (OT) systems coincides with this development. They described that models and techniques that take into account the functionality of the entire Cyber-Physical System are needed. To do this, a process-sensitive threat assessment for an attack response is combined with a context-based detection approach. Moreover, the adopted strategy must be flexible to consider the unpredictable nature of the method and the changing risks of cyberattacks. The results showed that cyber-attacks against cyber-physical systems (CPS) can be identified and differentiated using the technologies already in use. This suggested that monitoring the IT and OT components of the system is feasible to developing risk-based cybersecurity solutions. Tariq et al. [25] explained that for early recognition and preventing hostile activity and illegal access, intrusion detection systems (IDS) are essential to maintaining network security. Their study explained the design and evaluation of an IDS utilizing machine learning methods, with an emphasis on the oil and energy sector. Long short-term memory (LSTM), multilayer perception (MLP), random forest, and one-dimensional convolutional neural network (1DCNN) were among the models that were trained and tested using

both artificial and real-world datasets. With an area under the curve (AUC) of 96%, the 1DCNN model demonstrated the best performance among the models, demonstrating its efficacy in identifying network intrusions. Their study emphasized the importance of selecting the best machine learning algorithms for intrusion detection systems (IDS) and recommended further research into combination models and sophisticated architectures for advancements.

In another study, Pettersen and Grøtan [26] examined the oil and gas sector as an example of a field facing challenging circumstances and the immediate possibility of a major failure. The sector is going through major modifications due to advancements in digital technologies and is subjected to increasingly dangerous threats as a result of political shifts. It also involves cyber-physical systems, which have close connections between technological developments that can be triggered from almost anywhere. Using interviews, the study determined the degree to which the industry's current cybersecurity procedures could be improved by implementing resilience principles. The study highlighted the value of examining the empirical findings using a theoretical framework for assessing cyber resilience. Furthermore, they discovered that reducing the gap between strategic flexibility and cyber security resilience calls for a hybrid approach that combines robustness and adaptive capacity. Additionally, they found that a fundamental shift away from viewing resilience as a result and only an effect of current practice is necessary to prioritize adaptive capacity gradually. On the other hand, they viewed the adaptive capacity as resilience-as-process, a phenomenon that merits independent investigation. This suggests that managing cyber resilience needs to go beyond just integrating it with risk management.

Obonna et al. [27] highlighted that several amorphous cyberattacks have been launched on the oil and gas installation's process control network (PCN). Denial-of-service (DoS), distributed denial-of-service (DDoS), and man-in-the-middle (MitM) attacks are a few examples. One important influencing reason might have been the relatively inexpensive network development that led to the acceptance of public networks in operation technology (OT). The OT industry's connection to the internet for firmware updates, outside assistance, or supplier participation has made cyberattacks conceivable. These sporadic intrusions disclose the PCN when they go unnoticed, and an effective assault can have catastrophic results. In order to identify disparities, a study examined how machine learning techniques are used to monitor data exchanges among various network elements. It also reviewed the various forms of cyber-attacks in PCN of oil and gas installations. The experimental results demonstrated the accuracy and usefulness of various machine learning algorithms in identifying these anomalies, with notable precise attack detections identified using tree algorithms for man-in-the-middle (MitM) attacks accounting for trade-offs between precision and estimation complexity.

Shohoud [28] described that during the last ten years, Egypt's oil and gas sector's information technology usage has risen exponentially, along with the number of hacking attempts targeting these systems. The functioning and credibility of these businesses might be impacted greatly as a result of these threats to cyber security. Such attacks must be avoided at all costs. Particularly considering the importance of Egypt's oil and gas industry to the regional economy and the fact that a large number of these interconnected systems are occasionally controlled remotely. This study aimed to educate decision-makers on the significance of taking proactive steps to fortify the company's digital security and safeguard information-critical resources. It also analyzed the usefulness of the ISO 27001 standard [29] in reducing the impact of cyber threats. To achieve a high return on investment in understanding cybersecurity, the study highlighted the significance of enhancing the local educational system and applying an organized strategy that prioritizes behavioral change to close the gap between the supply and demand for cybersecurity specialists.

Progoulakis et al. [30] focused on enhancing the knowledge of dangers associated with cyber security and the administrative and technical protections that the oil and gas business must implement. The findings of the study discussed about cyber security for offshore oil and gas assets; providing insightful information about the mindset of the sector today and

how cyber security ideas are seen. The significance of business participation and support, employee involvement, training, organizational culture, and corporate support in the area of cyber security are emphasized. The study highlighted that the human aspects and the business structure must be used as two distinct perspectives to observe and comprehend the topic of cyber security. For the human elements, their survey's findings showed that threats from insiders and a lack of awareness of technology and culture are prioritized over most other cyber threats. Specifically, 73% of participants said it is likely that an insider will pose a threat to cyber security. Another real risk to cyber security breach situations is a lack of awareness of cyber-safety concepts and how they affect processes or a business in the event of an attack. Moreover, the results of the survey showed that, in addition to having disaster recovery strategies, oil and gas businesses either hire or receive assistance from outside cyber security specialists. Understandably, employees of oil and gas organizations had differing perspectives on the topic of cyber security when it comes to adaptability and comprehension.

Avanzini and Spessa [31] explained that combining old and modern technologies gives firms new perspectives and alternatives, but it also raises cybersecurity issues. Furthermore, due to their extreme visibility, critical sectors like oil and gas are increasingly vulnerable to cyberattacks. Their study offered a comprehensive cybersecurity strategy designed for the Oil and Gas (O&G) sector. The study addressed the three aspects of cybersecurity: people, processes, and technologies. For the asset owners, the process started with risk profile generation. High-risk items were ranked in order using a bowties and barrier management strategy. To identify the required mitigation actions security zones were established, Security Level Targets were created, and a gap analysis was carried out. In the next step of testing, phishing efforts and penetration tests were conducted. For manufacturers, the focus on the certification of the system and its components contributes to security. To ensure that cybersecurity resilience for O&G assets is accomplished, a comprehensive, methodology is intended to address all relevant factors from both an operational and organizational standpoint.

Mohammad et al. [32] stated that Supervisory Control and Data Acquisition (SCADA) is a critical component of ICPS as it provides process management and surveillance. These SCADA systems are known to interact using several insecure protocols that are open to different types of attacks. As a result, vital infrastructures, particularly those in the oil and gas industry, face higher cyber dangers. This study offered a method for attacking ICS to deny legal service using the Modbus TCP, in light of an increase in cyberattacks against these systems and the regularity with which these assaults result in DoS situations. This study presented a unique field flooding attack that can penetrate these defenses. The effects of the field flooding attack were assessed using three real industrial testbeds with various setups. The findings indicate that the programmable logic controller (PLC) often used in the oil and gas field is particularly susceptible to the assault, since a single erroneous packet caused a 59 min denial of service. In another study, Gueye et al. [33] highlighted the lack of real-world data needed to create neural network models efficiently, exploring the vital area of cybersecurity for industrial control and automation systems (ICS). This study aimed to fill the literature gap by assessing the effectiveness of a unique approach to ICS cybersecurity using data from real industrial settings. The study created a dataset using actual data from several commercial industries. These sectors include freshwater tanks, power networks, and gas pipelines. Authors claimed that the power system models obtain an astounding 71% accuracy rate, and the network performance is consistently increased by adding data produced. In several trials, the machine learning system achieved an astounding 99% accuracy using generated data. Furthermore, when the technique was used to set gas pipelines, most studies demonstrated that it was approximately 90% accurate. However, the study had some limitations, for example, because the study's restricted focus is on infrastructure, its findings cannot be easily applicable to other industries. The implementation of specific security methods against cyberattacks is a topic that the

present investigation did not include, indicating the necessity for additional studies on this important topic.

3.2. Cybersecurity Implications in Electricity

Electricity systems are a complex network spanning from energy generation plants, power grids, and distribution and transmission mechanisms, and recent cyberattacks advocate for a more serious focus on cybersecurity in the electricity domain [34]. There is a critical need for advanced cybersecurity measures to safeguard smart grids against vulnerabilities introduced by the integration of information and communication technologies (ICTs) [35,36]. Therefore, there is a need for collective efforts to strengthen global cybersecurity measures in the electricity sector by the development of strategic policies to safeguard energy infrastructure [37]. Similarly, Ratnam et al. [38] advocated for the resilience of the electricity systems as the complexity of modern-day electricity systems increased referred to as "Internet of Energy", so there is a need for improving the robustness of the electricity grids.

Sun et al. [39] applied intrusion detection mechanisms to improve the cybersecurity of the smart meters. The authors analyzed the vulnerabilities in the smart grids and proposed mechanisms to identify and avoid cyber intrusions. In another study, Shaaban et al. [40] adopted data-driven mechanisms for indicating electricity theft in photovoltaic generation systems. The study emphasized the importance of cybersecurity in making sure the integrity of photovoltaic (PV) generation, as electricity theft may compromise grid stability. Ibrahem et al. [41] highlighted how the variation and the transmit technique is performed to detect electricity theft at the Advanced Metering Infrastructure (AMI) systems. This study showed the elevating significance of securing AMI systems from cyber threats. In another study, Tolba and Al-Makhadmeh [42] proposed the utilization of an authentication approach with the goal of securing the communications in the smart grid environments. The results highlighted that this approach improved the cybersecurity of the grid communication systems by mitigating the main threats. A study by Johnson et al. [43] showed the cybersecurity obstacles for electric vehicles that were charged by this infrastructure. It showed the main threats and indicated the potential countermeasures to protect the charging networks from cyberattacks.

Furthermore, Bai et al. [44] introduced a model that aims to enhance the precision of power theft detection, by incorporating a transformer network with a Gaussian-weighted self-attention mechanism to capture global dependencies and temporal dependencies in the electricity consumption data. The framework addressed the impact of cyberattacks on critical infrastructure and discussed data preprocessing, normalization, and missing value processing. It also presented a neural network model and a subsystem for monitoring the network traffic. The proposed model was evaluated using two datasets, including the State Grid Corporation of China (SGCC) dataset, which was collected during the 2014–2016 period and is structured as time–series data, and another dataset obtained from the Canadian Institute for Cybersecurity. Musleh et al. [45] outlined that digital technologies improved the management and control of solar distributed generation systems; however, additional cybersecurity threats have emerged. The authors highlighted cybersecurity vulnerabilities of the distributed commercial solar inverters focusing on the Australian electricity grid. The authors experimentally showed the potential risks which affect grid stability; therefore, resilient cybersecurity measures are required. In another study, Erkek and Irmak [46] adopted the digital twin technology to improve the cybersecurity of the plant hydroelectric power. The authors highlighted how digital twin technologies may predict and model cyber threats, enhancing the plant's resilience against potential attacks. They applied the digital twin model to a power plant in Turkey to foster a proactive approach to cybersecurity.

Ismail et al. [47] adopted deep learning approaches to control electricity theft in the distributed generation (DG) domain. The main contribution of their work lies in addressing the manipulation of smart meters by malicious customers in renewable-based DG units to

overcharge utility companies. To reach this, the authors employed deep machine learning techniques, including deep feedforward, deep recurrent, and deep convolutional recurrent neural networks. The authors highlighted that smart meter data combined with meteorological data, and SCADA metering data can enhance the detection rate to 99.3% and false alarm to 0.22%. Similarly, Takiddin et al. [48] proposed the utilization of the variational auto-encoders for indicating stealth cyber-attacks at the advanced metering infrastructure networks. The approach implemented fully connected variational auto encoders and long and short-term memory variational auto encoders and was able to improve the detection rate in the range of 11–15%, 9–22% improvement in the false alarm rate, and 27% to 37% improvement was the highest difference compared to existing approaches. In another study, Takiddin et al. [49] developed a machine learning-based approach using vector embedding to detect electricity theft cyber-attacks at the AMI networks. This approach enhanced the detection effectivity and accuracy in indicating cyber threats in the systems of electricity distribution. The model was tested on two real datasets and achieved a 95.8% detection rate, 93.7% highest difference, and 2.1% false alarm. Tang et al. [50] provided an in-depth analysis of vulnerabilities in the demand-response systems integrating customer demand and smart grid response. In their experiment, a false demand is added to the system and an online detector using a convolutional neural network is made to control such demand requests. The system was trialed on an IEEE 34 bus feeder and the results highlighted that the developed system achieved higher accuracy and responded to cyber-attacks with fixed change rates. In another study, Heymann et al. [51] investigated the cybersecurity resilience in the Swiss electricity sector by researching 124 Swiss energy market representatives. The research study highlighted policy recommendations to improve the energy sector's protection against cyber threats and improved the system's stability. The study advocated for stringent regulatory measures and monitoring strategies to improve the Swiss energy sector's cybersecurity resilience. Another study [52] pointed out the frequency of the synchronization consensus issue in networked microgrids vulnerable to multi-layer denial of service (DoS) attacks, which may concurrently impact measurements, control activation, and communication pathways. A unified concept called Persistency-of-Data-Flow (PoDF) was put forth to quantify the multi-layer DoS effects on the hierarchical system and characterize the data unavailability in various information network linkages. They provided a condition of DoS attacks with PoDF that enabled consensus maintenance of the proposed edge-based self-triggered distributed control system. To mitigate the conservativeness of offline design against the worst-case assault across every device, an online self-adaptive strategy of the control parameters was also built to fully exploit the most recent information of all data transmission channels. Lastly, illustrative case studies were used to confirm the efficacy of the suggested cyber-resilient self-triggered distributed control.

In the literature, some studies carried out detailed reviews on the cybersecurity implications specifically in the electricity domain. For instance, Liu et al. [53] highlighted the use of digitally controlled and software-driven distributed energy resources (DERs) to enhance grid operations. However, this development also makes geographically scattered DERs vulnerable to digital threats, such as staff mistakes, communication problems, and hardware and software flaws which enforces the importance of cybersecurity in this area. In this regard, they have given a detailed overview of the advancements in cyber-resiliency enhancement (CRE) of the DER-based smart grid. Firstly, a holistic threat modeling approach with a focus on effect analysis and identifying vulnerabilities was specifically designed for the hierarchical DER-based smart grid. The defense-in-depth tactics that include detection, avoidance, mitigation, and restoration are then thoroughly examined, categorized, and meticulously compiled. The five main resiliency enablers were then incorporated into a comprehensive CRE framework. Lastly, a thorough discussion of the difficulties and potential paths forward was provided. Similarly, Nafees et al. [54] emphasized that opponents can launch sophisticated cyberattacks, including advanced persistent threats and coordinated attacks, resulting in operational issues and, in the worst cases, blackouts of electricity as a result of the substantial rewards that grid dangers can understand. The

Ukrainian power grid attack exemplified the latter. In their study, they examined the nature of cyber-physical threats to comprehend their features and provided a threat modeling methodology. In particular, they explored the nature of cyber-physical threats and provided a threat modeling framework to comprehend their traits and effects on the physical and control systems of the smart grid. They also looked at current threat detection and defense capabilities. Moreover, they explained how electricity system managers should include human factors while assessing the effects of intrusions. Zhang et al. [55] highlighted the application of machine learning (ML) on Internet of Things (IoT)-based smart grids. They emphasized that the usual management and operation of the equipment will be significantly impacted by the hostile disturbance introduced into the power stream. As a result, security evaluation in safety-critical power systems is essential. They thoroughly analyzed the latest developments in attack and defense strategy design for ML-based smart grids. The study drew attention to the details involved in creating hostile attacks against these ML-based smart grids. They carried out a thorough investigation to examine and contrast previous research on adversarial assaults on Machine Learning-based smart grids in situations including initiation, broadcast, supply, and utilization. The countermeasures were then evaluated based on the attacks they were designed to fend off. Lastly, the attacker's and defender's respective future research directions were examined. In another study, Inayat et al. [56] highlighted that the smart grid needs to be protected from growing security risks and intrusions. In their study, they presented a thorough analysis of protection strategies that can be applied to identify these kinds of assaults and lessen the risks they pose. They emphasized that to reduce the frequency of cyberattacks, the targeted equipment's security needs to be improved by mentioning different methods and models such as the Gaussian process model, pattern detection method, honeypots, parametric feedback linearization controller, etc.

3.3. Cybersecurity Implications in Nuclear Energy

Due to the digitalization of control systems, cybersecurity in digital control systems is very important specifically focusing on vulnerabilities in the network infrastructure and control systems as well as human aspects [57]. There is a need to adopt a three-pronged policy focusing on enhancing cybersecurity resilience, foster public–private partnerships for cyber-attack preparation, and improving the security of nuclear systems [58]. Similar concerns were highlighted by Falowo et al. [59] that the need for clean energy sources has triggered increasing demand for nuclear fusion plants, so there is a critical need for cybersecurity management of such infrastructures. They highlighted that startup companies due to their increased innovation might help in bringing cybersecurity agility in protecting nuclear fusion plants, where threats may be classified based on the importance of the infrastructure and the nature of the threat for continuous threat assessment.

Jung et al. [60] developed a technical assessment methodology to evaluate potential attacks on an asset and applicable controls. They compared their assessment result by jointly implementing their assessment along with NEI 13-10 (cybersecurity control assessments) against the application of on NEI 13-10 on plant protection of a nuclear power reactor APR 1400. The study provided insights on further improving the assessment mechanisms.

Yockey et al. [61] highlighted it is critical that system designers, reactor operators, and regulators must focus on the cybersecurity implications during the design of autonomous control systems (ACS) in advanced nuclear reactors. The study developed a cyber-physical testbed using digital twin technologies. The testbed included two plant-level digital twins and two component-level digital twins for reactor malfunction/control action and component states/forecasting component input/output, respectively. Furthermore, two duplicate ACS designs were formulated one based on traditional machine learning and second, an automated machine learning approach, and the results highlighted that neither of them is optimal. Hence, the authors laid out a set of recommendations to all stakeholders to foster a shared responsibility for securing machine learning-based systems.

Table 1 provides a summarized contribution of all papers included in this review.

Table 1. Overview of Studies included in the Review.

S. No	Publication Year	Main Contribution	Technology/Method
[23]	2023	**Context and Problem:** Highlighted that the global cyberattacks increased by 50% in 2021 as a result of the pandemic's intensification in 2020. This scenario has been made worse by the conflict in Ukraine since 2022, especially in the oil and gas sector, which is considered essential infrastructure, confronts cybersecurity issues that necessitate a proactive strategy that combines technology and behavioral controls. **Solution and Result:** Proposed a cybersecurity framework model that was assessed through implementation control and staff survey and showed 92.69% efficacy and 81.55% acceptance by staff.	Derived from the Center for Internet Security Critical Security Controls (CIS CSC) and the National Institute of Standards and Technology Cybersecurity Framework (NIST CSF).
[24]	2023	**Context and Problem:** Emphasized that the industrial control systems (ICS) and industrial automation and control systems, which have traditionally been protected from cyberspace, are at risk of cyberattack. They highlighted attacks such as the US Colonial Pipeline attack, Ukraine Grid Attack, and Norway Oil Platform attack. **Solution and Results:** Proposed a context-based detection approach integrated with a knowledge-based approach to mitigate the effect of cyberattacks. Existing monitoring applications can be utilized to identify and differentiate between various cyberattack types. This shows that it is feasible to monitor the ICS system's IT and control components in order to develop risk-based cybersecurity decision support systems.	Integrating, a process-sensitive threat assessment for attack response with a context-based detection approach.
[25]	2023	**Context and Problem:** Emphasized the significance of selecting suitable machine learning algorithms for intrusion detection systems (IDS) in the oil and gas industry. Four machine learning algorithms were evaluated in this context. **Solution and Results:** Highlighted that the 1DCNN model achieved the highest performance with 96% accuracy.	Machine learning algorithms for intrusion detection in oil and gas industry using intrusion detection dataset.
[26]	2024	**Context and Problem:** Mentioned that the oil and gas industry is vulnerable to cyberattacks because of the digital transformation. They explored the degree of resilience by evaluating the oil and gas industry's current cybersecurity procedures. **Solution and Results:** Examined the empirical data by proposing a "resilience ABC" which takes into account a significant difference between resilience based on adaptive capacity and robustness.	Empirical study.
[27]	2023	**Context and Problem:** Highlighted the dangers of cyber-attack (such as DoS, DDoS and MitM) on process control network (PCN) of the oil and gas industry. The PCN is exposed by its incapacity to identify these dangerous cyberattacks, and a successful attack could have disastrous consequences. **Solution and Results**: Performance evaluation of various machine learning techniques for detection of MitM attacks in a process control network in an oil and gas installation. Coarse tree algorithm showed high performance for identifying the MitM attack.	Machine learning techniques for detection of MitM attacks using real time dataset.
[28]	2023	**Context and Problem:** Underscored that the Egyptian oil and gas industry have gone through a digital transformation which led to several security breaches. **Solution and Results:** Investigated the benefits of implementing ISO 27001 for reducing cyber threats in Egypt's downstream oil and gas industry and also raised cybersecurity awareness in the oil and gas industry.	Empirical study.

Table 1. *Cont.*

S. No	Publication Year	Main Contribution	Technology/Method
[30]	2021	**Context and Problem:** Highlighted the offshore oil and gas industry is facing cyber threats because of digitization. **Solution and Results:** Explored the risks to cyber security through a survey study and recommended organizational (such as cybersecurity awareness and training) and technical safeguards (such as real-time monitoring) that the oil and gas sector should implement.	Empirical study/survey.
[31]	2019	**Context and Problem:** The oil and Gas industry is vulnerable to various high-profile cyberattacks because of its critical infrastructure. This may lead to heavy economic damage as well as a threat to people's security and the environment. **Solution and Results:** It outlines a comprehensive strategy to cybersecurity designed specifically for the oil and gas industry. This strategy addresses issues concerning technologies, people, and procedures, or the "three pillars" of cybersecurity. Moreover, devised cybersecurity strategy guidelines by integrating operational and organizational standpoint.	Proposed holistic framework and recommendations for cybersecurity resilience.
[32]	2023	**Context and Problem:** The oil and gas industry rely heavily on SCADA system which uses insecure communication protocols. It leads to several cyberattacks such as DoS. **Solution and Results:** Presented a unique field flooding attack by conducting an experimental study and highlighted that the PLC often used in the oil and gas field are particularly susceptible, since a single erroneous packet caused a 59 min denial of service. This algorithm showed 99% accuracy.	Evaluation of Field flooding attack on the network based on 4 h of network capture traffic from three testbeds to formulae dataset.
[33]	2024	**Context and Problem:** Cyberattacks can destroy and damage critical infrastructures such as power, water, and gas because of the lack of real-world industrial control and automation systems. **Solution and Results:** Assessed the effectiveness of cybersecurity techniques used in industrial control systems using real-time data and formed a combined dataset. Results showed that the dataset quality affects the model's performance.	Machine learning applications on three datasets of power system, freshwater tank, and gas pipeline.
[39]	2020	**Context and Problem:** Advanced metering infrastructure is vulnerable to cyberattacks because of digitization and can affect consumers. **Solution and Results**: Introduced two-stage intrusion detection mechanisms for the cybersecurity of smart meters in power grids which effectively identified the cyberattacks in smart meters.	Two staged intrusion detection for smart meters.
[40]	2021	**Context and Problem:** Electricity units are at risk of cyberattack by malicious consumers who may change their data reading in smart meters leading to electricity theft. **Solution and Results:** Focused on detecting electricity theft in photovoltaic (PV) generation using a data-driven method based on a regression tree. Performance of regression tree is compared with other models which showed better performance by regression tree.	Data-driven detection of cyberattacks in PV generation
[41]	2022	**Context and Problem:** The change and transit approach is very commonly used in smart metering systems, but it has brought challenges of vulnerability to cyberattacks which can lead to electricity theft, financial loss, and grid instability. **Solution and Results:** Deep learning-based solutions for detecting electricity theft in Advanced Metering Infrastructure (AMI) systems which outperform the traditional methods.	Deep learning-based change and transmit detection techniques in AMI networks

Table 1. *Cont.*

S. No	Publication Year	Main Contribution	Technology/Method
[42]	2021	**Context and Problem:** Common challenges in digitizing the power grid include security threats such as false data injection, which diminish the predicted assimilation performance. **Solution and Results:** Presented a user authentication approach to secure smart grid communications which improves the detection of false data injection more effectively.	Cybersecurity user authentication for smart grids.
[43]	2022	**Context and Problem:** Electric vehicle chargers when interacting with grid stations pose several cybersecurity vulnerabilities that can lead to financial loss and grid instability. **Solution and Results:** Analyzed cybersecurity threats related to electric vehicle charging infrastructure and proposed measures for securing EV chargers from attack.	Cybersecurity measures for electric vehicle charging infrastructure.
[44]	2023	**Context and Problem:** Electricity theft is a major factor in power outages. In recent years, there has been rising recognition of using neural network models in electrical theft detection (ETD). However, conventional techniques have a limited ability to gather deep properties, making it difficult to spot abnormalities in power consumption data consistently. **Solution and Results:** A model that aimed to enhance the precision of power theft detection using a transformer network with a Gaussian-weighted self-attention mechanism to capture global and temporal dependencies in electricity consumption data.	An experimental study using two datasets, including the State Grid Corporation of China (SGCC) and another dataset obtained from the Canadian Institute for Cybersecurity.
[45]	2024	**Context and Problem:** Combining solar distributed generation (DG) devices into the electricity grid adds complexity that might affect the grid's dependability and security. **Solution and Results:** Evaluation of cybersecurity vulnerabilities and impacts of distributed solar inverters on the Australian grid.	Experimental evidence of cybersecurity vulnerabilities of distributed commercial solar inverters
[46]	2024	**Context and Problem:** Hydroelectric power plants face cyberattacks because they integrate into digital systems. **Solution and Results:** Enhancing cybersecurity of a hydroelectric power plant in Turkey using a digital twin model to detect and analyze attacks. Results showed that it improves threat detection.	Digital twin model.
[47]	2020	**Context and Problem:** The distributed generation domain is vulnerable to attack as malicious user can change the meeting readings, leading to electricity theft. **Solution and Results:** Developed a deep learning–based system to detect electricity theft in renewable distributed generation (DG) using novel cyber-attack functions. The model has the highest detection rate (99.3%) and the fewest false alarms (0.22%).	Utilized deep feed forward, deep recurrent, and deep convolutional recurrent neural networks for detection. Created datasets from smart meter readings, meteorological (solar irradiance) data, and SCADA metering data, simulating an IEEE 123 bus test system.
[48]	2020	**Context and Problem:** AMI networks face cyberattacks because of the malicious data given to them. Traditional models are unable to deal with this issue and are unable to detect electricity theft. **Solution and Results:** Detection of stealth cyber-attacks in AMI networks using variational auto-encoder-based techniques. Improve the detection rate by 11–15%, false alarm rate by 9–22%, and highest difference by 27–37% over existing detectors.	Variational auto-encoder.

Table 1. *Cont.*

S. No	Publication Year	Main Contribution	Technology/Method
[49]	2020a	**Context and Problem:** Electricity theft is difficult to identify because of false energy consumption data and the legacy ML models are unable to identify these thefts. **Solution and Results:** Detection of electricity theft cyber-attacks in AMI networks using deep vector embeddings. The proposed model outperforms the shallow detectors showing high performance and accuracy.	Deep vector embeddings.
[50]	2023	**Context and Problem:** Integration of demand response programs in smart grids poses cyber security threats due to false data injection. **Solution and Results:** Explored vulnerabilities in demand-response systems with renewable energy integration under cyberattacks and proposed an online detector for cyberattacks. Results showed that detectors helped in effectively mitigating the attacks.	Vulnerability analysis of demand-response in smart grids.
[51]	2022	**Context and Problem:** The Swiss electricity system is prone to cyberattacks because of digital transformation. **Solution and Results:** Cybersecurity and resilience measures in the Swiss electricity sector, offering policy options for enhancement, which showed that the cybersecurity system needs improvement.	Participant feedback, cybersecurity, and resilience analysis.
[52]	2023	**Context and Problem:** Network microgrids face cyberattacks, especially from multi-layer DoS attacks. **Solution and Results:** Construct an online self-adaptive strategy of the control parameters to fully use the most recent information of all data transmission channels, hence mitigating the conservativeness of offline design against the worst-case attack across all devices.	Cyber-resilient self-triggered distributed control to mitigate multi-layer DoS attacks.
[60]	2023	**Context and Problem:** The nuclear industry is introduced to cybersecurity attacks because of digitization. **Solution and Results:** Devised a methodology for cybersecurity controls assessment of nuclear powerplant which offers a comprehensive understanding of cyberattacks.	Cybersecurity assessment framework.
[61]	2023	**Context and Problem:** Nuclear power plants and energy plants are becoming vulnerable to cyberattacks **Solution and Results:** Developed a cyber-physical testbed using digital twin technologies. The testbed included two plant-level and digital twins and two component-level digital twins for reactor malfunction/control action and component states/forecasting component input/output, respectively.	Digital twin, machine learning

4. Discussion

Cybersecurity implications of the digital transformation in the energy sector are very critical due to the enhanced importance of energy assets [3,4]. A taxonomy of cybersecurity in energy sector organizations is developed and is shown in Figure 3. The energy sector can be categorized into three sub-sectors: the oil and gas sector, the electricity and renewable energy sector, and the nuclear energy sector. Power plants, oil and gas pipelines, smart distribution grids, energy storage systems and offshore platforms are key infrastructure components of energy sector organizations that need to be secured. The sub-components in these organizations can be categorized into control systems, operational technologies, information technology tools and integrated information, and operational technology systems. Control systems may include supervisory control and data acquisition (SCADA) [21,32,47], distributed control systems (DCS) [62], or programmable logic controllers (PLCs) [32]. The security of these control systems is important as attacks on these control systems will result in the loss of control of these critical infrastructures. Cybersecurity attacks on operational technologies like real-time systems such as power generation controls [63], grid monitoring controls, turbine controls [64], and battery optimization systems can result in outages and

malfunction of energy sources. Information technology infrastructure components such as networks, data centers, and cloud systems are another key target of cybersecurity attackers to sabotage the critical infrastructures. The last sub-component integrates the information technology components with operational technologies to enhance the performance of energy sector organizations. It may include simulation and modeling tools of OT data [65], grid analytics [66], smart meters, and battery analytics [67]. Cybersecurity attacks on these components may result in significant performance degradation and service delays.

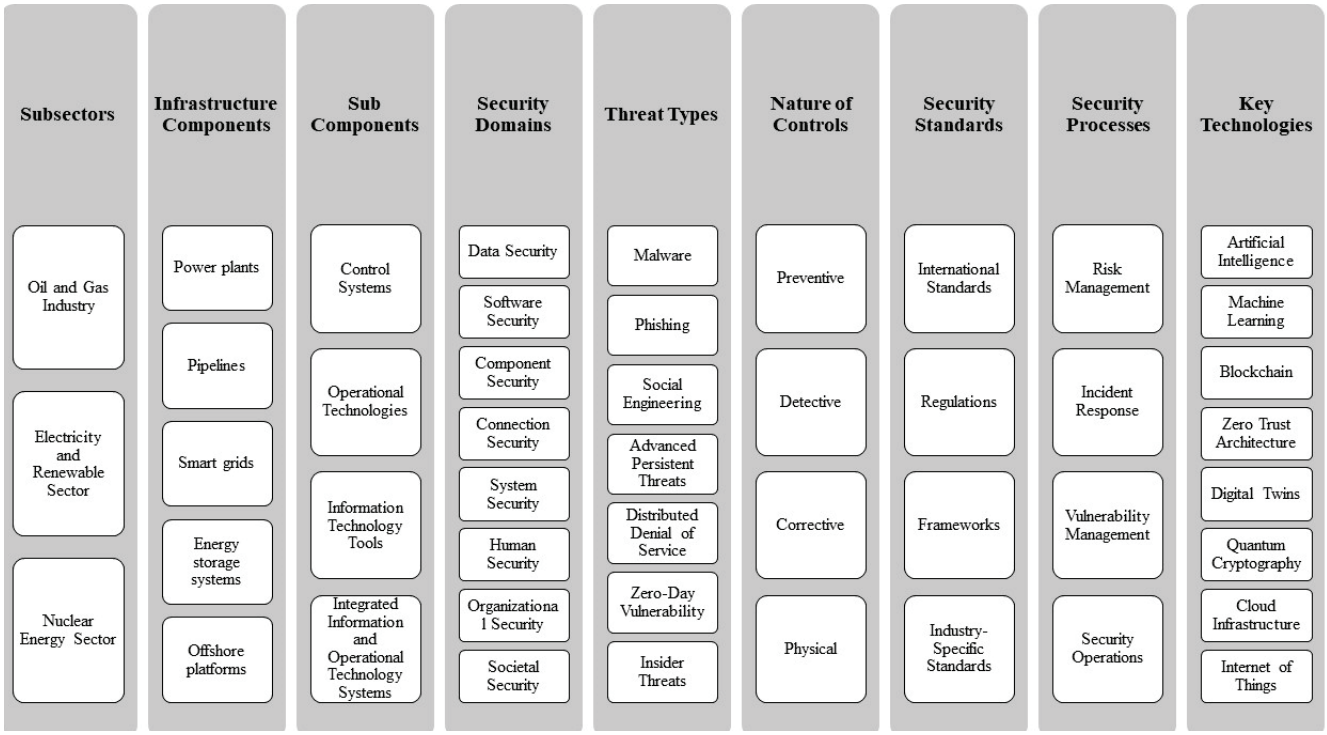

Figure 3. Taxonomy of Cybersecurity Implications in the Energy Sector.

The Institute of Electrical and Electronics Engineers [68] and Association of Computing Machinery have [69] classified the security in eight domains and all these domains are relevant to cybersecurity in energy sector organizations. The first domain is human security which refers to securing individuals, and data, and analyzing human behavior for cybersecurity. Energy sector organizations can develop cybersecurity measures in place which can enhance social engineering cybersecurity, by focusing on human factors, identity management, and increasing cybersecurity awareness and usable privacy and security among all stakeholders [9]. The data security aspect investigates securing the data in the organization and energy sector organizations can foster cybersecurity controls such as quantum cryptography, digital forensics, access control, data integrity, data storage security, and secure communication protocols. Software security refers to the adaptation of cybersecurity principles during the software development and usage stage. Secure software design, usability, and rigorous testing are key aspects in this domain and it can be specifically challenging when software development is carried out by supply chain partners of energy sector organizations. Approaches like zero trust architecture [70] and DevSecOps [71] can improve software security. Component security refers to cybersecurity during the design, acquisition, testing, and deployment phases of components in larger systems. Component reverse engineering [72], design, testing, and procurement processes in energy sector organizations need to be analyzed to enhance component security. Furthermore, connection security refers to cybersecurity challenges related to the establishment of connections between different organization infrastructure components. In this domain cybersecurity challenges pertaining to network architecture, hardware architecture, dis-

tributed systems architecture, network services, and defense are key challenges. The system security domain looks at the security challenges from a holistic view of systems integration and challenges like system management, system access, and control and system retirement are taken into consideration. The organizational security domain focuses on protecting the organization from cybersecurity attacks and challenges like risk management [26], security governance, cybersecurity planning, security analytics, and security operations that fall under this domain. Lastly, the societal security domain refers to cybersecurity issues that impact society, and areas like cybercrimes, cyber laws, and ethics are key issues in this domain. Energy sector organizations can focus on satisfying regulatory and legal requirements [51] pertaining to the cybersecurity of their organizational infrastructure.

As shown in Figure 3, threats can be classified as malware [73], phishing attacks [74], social engineering attacks [75], advanced persistent threats (APTs) [76], DDoS attacks [27], zero-day vulnerabilities [77], and insider threats [78]. An extensive security policy capable of resilience against these diverse cybersecurity attacks can improve the security resilience of energy sector organizations. A balanced combination of preventive, detective, corrective, and physical controls in energy sector organizations can help in achieving cybersecurity goals. Additionally, organizations need to implement international standards (such as NIST 800-53 [79], ISO/IEC 27001 [80], IEC 62443 [81]), regulations (such as General Data Protection Regulation (GDPR) [82]), frameworks (such as Network and Information Systems directive), and industry-specific standards (NIS 2 [83], Presidential Policy Directive (PPD) [84]) to make their cybersecurity controls more effective. Cybersecurity processes in energy sector organizations can be categorized into risk management, incident response, vulnerability management, and security operations where advanced technologies such as Artificial intelligence, machine learning [25,27,33], blockchain [85], quantum computing [86], zero trust architecture [70], cloud computing [87], digital twins [46], and internet of things [22,55] can support in establishing a robust cybersecurity framework.

As shown in Figure 4, we present a list of typical challenges faced by energy sector organizations to protect the energy infrastructure from cybersecurity attacks. Based on this we present a set of key research directions which can help other researchers to work on improving the cybersecurity of energy sector infrastructures.

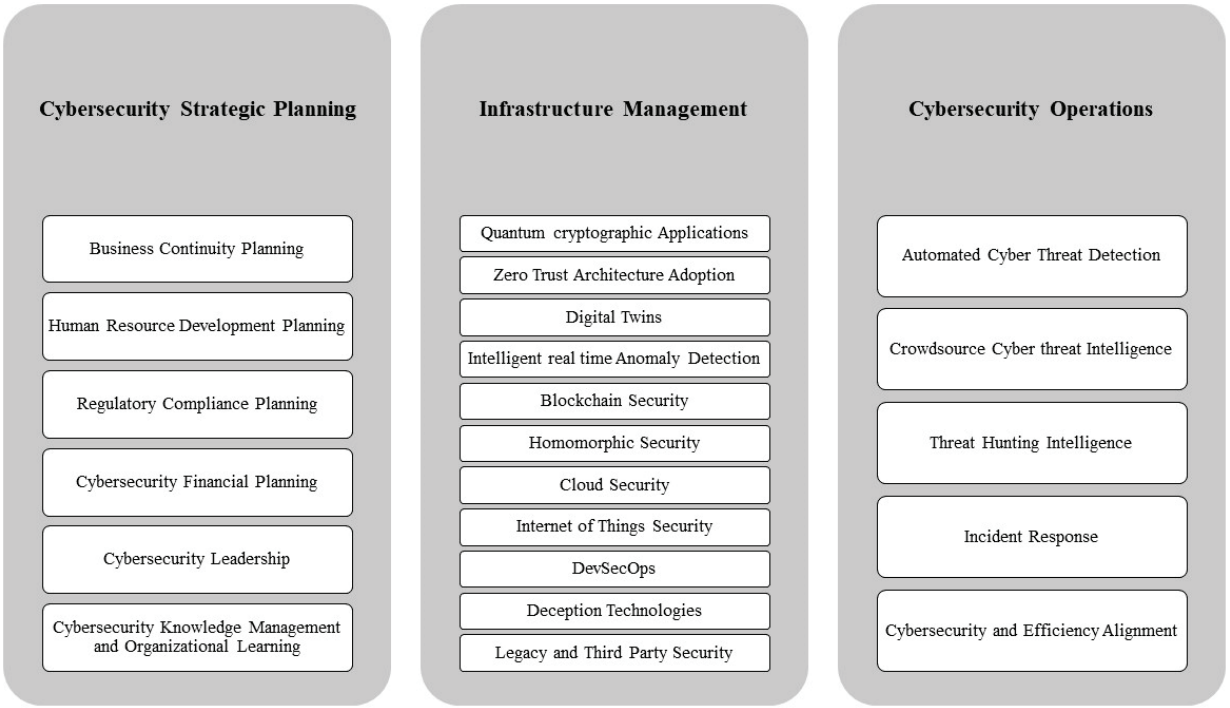

Figure 4. Cybersecurity Challenges in the Energy Sector.

Cybersecurity strategic planning ensures that organizations adopt a systematic approach to protect their organization from cybersecurity threats. Energy sector organizations being the critical infrastructure need to have a robust business continuity plan. Business continuity planning ensures that organizations can recover from cybersecurity attacks and resume operations quickly. A robust risk management process needs to be established to identify critical assets, identify threats and their impact. After this, the recovery time objectives and resources need to be allocated which can develop appropriate controls to enhance the business resilience. Additionally, an incident response plan needs to be developed which outlines the roles and responsibilities of response teams, protocols of response reporting, and a communication plan. The communication plan should include the strategy of communication with all stakeholders and regulatory authorities during the cybersecurity attacks. The plan should also define the schedule of data backup, regular recovery testing trials, and backup infrastructure. Successful and failed case studies of business continuity cybersecurity planning in energy organizations can facilitate other organizations in designing their cybersecurity policies. Typical research in diverse organizational and cultural contexts focuses on questions like—How warfare, international politics, and business competition can contribute in cybersecurity risks? How cybersecurity resilience can be enhanced in energy sector by optimizing business continuity planning? What are effective communication protocols in cybersecurity response? What are best practices in cybersecurity recovery? —can strengthen this body of knowledge.

Human resources are key in improving cybersecurity practices and energy sector organizations can develop a systematic approach to human resource development. Such an approach should include hiring key talent and designing skill development programs to foster a cybersecurity culture within the organizations. Additionally, comprehensive cybersecurity awareness plans need to be designed. The plan may include simulated social engineering and cyber-attack scenarios for employees to become acquainted with cybersecurity threats. Gamification and storytelling approaches can be integrated to enhance the user experience [88]. Here, research questions like—How do cultural and organizational challenges affect the stakeholder's cybersecurity readiness? How stakeholders' cybersecurity awareness can be enhanced? How does gamification help in improving employee cybersecurity readiness? Empirical studies answering such questions can contribute to fostering cybersecurity resilience in energy sector organizations? —can contribute to literature. It is also critical to educate users about security implications to improve external and internal threat management within the energy sector organizations [89,90]. There have been some studies showing that employees have limited information security readiness [91], so cybersecurity training [92,93] can better prepare them to cope with the cybersecurity challenges and optimally respond to threat situations. Such a rich research agenda can help the energy sector organizations to secure their digital transformation drive by adopting effective controls and risk management methodologies [94–97].

Organizations need to align their cybersecurity practices with government cybersecurity laws and standards to proactively respond to cybersecurity challenges and enhance stakeholders' trust, therefore energy organizations need to do comprehensive planning. Empirical research can enhance this body of knowledge by documenting best practices and challenges for energy organizations in adhering to standards and improving the planning process. Additionally, researchers can investigate on developing new standards and frameworks to enhance cybersecurity in energy sector organizations. Furthermore, cyber laws and policies can be investigated for cybersecurity enhancements in the energy sector supply chain. Financial implications of cybersecurity attacks are also very critical and organizations need to plan the budget allocation for cybersecurity compliance, emergency risks, as well as insurance. Researchers can explore questions like—How can cybersecurity financial forecasting be enhanced? How can critical infrastructure cybersecurity operational cost be minimized? How can models be optimized to predict the financial value of critical assets for insurance?

Cybersecurity leaders need to drive the cybersecurity processes, so they need to hold the appropriate skills to excel. Additionally, cybersecurity is a knowledge-intensive activity so cyber leadership in organizations needs to plan the fostering of effective knowledge management processes and tools to continuously learn and improve cybersecurity resilience. In this context, typical research questions such as—How can the cyber leadership process be sustained in energy sector organizations? How can skills be enhanced of cybersecurity leaders? How can culture influence the cybersecurity behavior of leadership? How can organizational learning be fostered in cybersecurity activities of energy sector organizations? How can appropriate tools be designed to support knowledge management in cybersecurity operations of energy sector organizations?—maybe considered. It is also very interesting to document in-depth case studies about leadership roles during the cybersecurity response to document best practices [98,99].

Technological infrastructure is another critical factor that focuses on aligning appropriate advanced technologies with organizational processes to gain optimal advantages. Cybersecurity activities are heavily reliant on technologies and energy sector organizations need to enhance the technology management processes. Firstly, as we have seen some studies have used machine learning [27,33,61], and digital twin [46,61] technologies to improve their critical infrastructures from cybersecurity threats so we highlight that there is a need for more studies to design secure technologies using advanced technologies [73]. Furthermore, the adoption of user-centric design [100] approach while designing security control systems can help to foster better adoption of security controls. The research community can develop advanced technologies to improve cybersecurity processes. The following research questions may be answered: How can quantum cryptography enhance the resilience of cryptographic algorithms? How can digital twin technologies and zero trust architecture be used to protect digital assets in energy sector organizations? What are challenges in fostering such technologies in energy sector organizations? How machine learning and deep learning models can be enhanced for intelligent anomaly detection? What are the key operational issues in adopting blockchain security, homomorphic security, cloud, and internet of things security? How DevSecOps approach can be fostered in the software development process of energy sector applications? What kind of deception technologies can help in minimizing the cybersecurity risks for energy sector organizations? How security of legacy systems and third-party components can be enhanced? Additionally, case studies of best practices in technology adoption in cybersecurity can be documented by conducting empirical research.

Cybersecurity operations is the process that scans the threats in the organizational environment and responds. Threat identification and response are the key activities in this stage. Adoption of intelligent tools and crowdsourcing for threat hunting enhances the effectiveness and efficiency of cyber operations. Furthermore, effective incident response mechanisms can help in minimizing downtime of critical infrastructures. A consistent challenge operation teams need to face is to find a balance between cybersecurity and the efficiency of infrastructure. How can researchers investigate how energy sector organizations can benefit from advanced cyber threat intelligence tools and techniques? How can incident response mechanisms in critical infrastructure be enhanced? In depth case studies of incident response and cyber threat intelligence can help cybersecurity teams in improving their security operations in energy sector organizations. Furthermore, there is a need for in-depth empirical studies of cyber threat intelligence of operations in energy companies, oil and gas companies, as well as nuclear plants can further enhance the understanding of threat identification, threat management, and threat mitigation challenges faced by the organizations.

5. Conclusions

The energy sector is a critical infrastructure for any nation and the digital transformation drive in this sector is critical for achieving efficiency. However, such a digital transformation drive has resulted in massive cybersecurity challenges. In this paper, we

have carried out a systematic literature review on cybersecurity implications in the energy sector and based on our review we highlight that there is a need to enrich this body of knowledge by improving security controls and technologies, a rich set of empirical studies documenting in-depth analysis of operational cybersecurity responses and user cybersecurity training. Our findings will help research teams to explore the research agenda to explore and improve cybersecurity in this important sector.

Author Contributions: Conceptualization, S.S.; methodology, S.S. and H.G.; data curation, A.F.A., M.S.A. and M.M.A.; writing—original S.S., M.S.A., A.F.A., M.M.A., H.G., M.S. and S.Z.I.; draft preparation, M.S.; writing—review and editing, S.S.; funding acquisition, A.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank Saudi Aramco Cybersecurity Chair for supporting this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. IEA. 2024. Available online: https://www.iea.org/reports/electricity-2024/executive-summary (accessed on 25 September 2024).
2. Xu, Q.; Zhong, M.; Li, X. How does digitalization affect energy? International evidence. *Energy Econ.* **2022**, *107*, 105879. [CrossRef]
3. Maroufkhani, P.; Desouza, K.C.; Perrons, R.K.; Iranmanesh, M. Digital transformation in the resource and energy sectors: A systematic review. *Resour. Policy* **2022**, *76*, 102622. [CrossRef]
4. Akberdina, V.; Osmonova, A. Digital transformation of energy sector companies. *E3S Web Conf.* **2021**, *250*, 06001. [CrossRef]
5. Nazari, Z.; Musilek, P. Impact of digital transformation on the energy sector: A review. *Algorithms* **2023**, *16*, 211. [CrossRef]
6. Oudina, Z.; Derdour, M.; Dib, A.; Yaakoubi, M.A. Identifying and Addressing Trust Concerns in Cyber-Physical Systems for the Oil and Gas Industry. *Ing. Syst. D'inform.* **2024**, *29*, 469–478. [CrossRef]
7. Gutman, S.; Brazovskaia, V. Tool Development for Assessing the Strategic Development of Territorial Socio-Economic Systems for the Purposes of Energy Sector Digital Transformation. *Energies* **2023**, *16*, 5269. [CrossRef]
8. Saeed, S.; Altamimi, S.A.; Alkayyal, N.A.; Alshehri, E.; Alabbad, D.A. Digital transformation and cybersecurity challenges for businesses resilience: Issues and recommendations. *Sensors* **2023**, *23*, 6666. [CrossRef]
9. Saeed, S. Usable Privacy and Security in Mobile Applications: Perception of Mobile End Users in Saudi Arabia. *Big Data Cogn. Comput.* **2024**, *8*, 162. [CrossRef]
10. Gull, H.; Saeed, S.; Alaied, H.A.; Alajmi, A.N.; Saqib, M.; Iqbal, S.Z.; Almuhaideb, A.M. Digital Transformation of Marketing Processes, Customer Privacy, Data Security, and Emerging Challenges in Fostering Sustainable Digital Marketing. In *Ethical AI and Data Management Strategies in Marketing*; Saluja, S., Nayyar, V., Rojhe, K., Sharma, S., Eds.; IGI Global Scientific Publishing: Hershey, PA, USA, 2024; pp. 71–88. [CrossRef]
11. Langner, R. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Secur. Priv.* **2011**, *9*, 49–51. [CrossRef]
12. Hobbs, A. *The Colonial Pipeline Hack: Exposing Vulnerabilities in US Cybersecurity*; SAGE Publications: SAGE Business Cases Originals: London, UK, 2021.
13. Cunningham, C. *A Russian Federation Information Warfare Primer*; The Henry M. Jackson School of International Studies, Washington University: Seattle, WA, USA, 2020.
14. Alqurashi, R.K.; AlZain, M.A.; Soh, B.; Masud, M.; Al-Amri, J. Cyber attacks and impacts: A case study in Saudi Arabia. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 217–224. [CrossRef]
15. Bhattacharjee, S.; Das, S.K. Detection and forensics against stealthy data falsification in smart metering infrastructure. *IEEE Trans. Dependable Secur. Comput.* **2018**, *18*, 356–371. [CrossRef]
16. Oudina, Z.; Dib, A.; Yakoubi, M.A.; Derdour, M. Comprehensive Risk Classification and Mitigation in the Petroleum Cyber-Physical Systems of the Oil and Gas Industry. *Int. J. Saf. Secur. Eng.* **2024**, *14*, 99–113. [CrossRef]
17. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
18. Google Scholar. Available online: https://scholar.google.com/schhp?hl=en (accessed on 26 September 2024).
19. Aubuchon, T.; Susanto, I.; Peterson, B.T. Oil and Gas Industry Partnership with Government to Improve Cybersecurity. In Proceedings of the SPE International Oil and Gas Conference and Exhibition in China, Beijing, China, 5–7 December 2006; p. SPE-104284.

20. Goel, A. Cybersecurity in O&G Industry. In Proceedings of the Offshore Technology Conference, Houston, TX, USA, 1–4 May 2017; pp. 6–9.
21. Mohammed, A.S.; Reinecke, P.; Burnap, P.; Rana, O.; Anthi, E. Cybersecurity challenges in the offshore oil and gas industry: An industrial cyber-physical systems (ICPS) perspective. *ACM Trans. Cyber-Phys. Syst. (TCPS)* **2022**, *6*, 28. [CrossRef]
22. Frederick, B.A.; Taylor, O.E. Analysis on Cybersecurity Control and Monitoring Techniques in Industrial IoT: Industrial Control Systems. *Internet Things and Cloud Comput.* **2023**, *11*, 1–17.
23. Villarreal, R.; Alarcón, F.; Torrejón, L. Implementing NIST CSF and CIS CSC in Gas Industry: A Model's Effectiveness and Acceptance Analysis. In Proceedings of the 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 16–17 November 2023; pp. 1–6.
24. Houmb, S.H.; Iversen, F.; Ewald, R.; Færaas, E. Intelligent risk based cybersecurity protection for industrial systems control—A feasibility study. In Proceedings of the International Petroleum Technology Conference, Bangkok, Thailand, 1–3 March 2023; p. D021S014R001.
25. Tariq, A.; Elhadef, M.; Ghani Khan, M.U. Optimizing Cybersecurity in the Oil and Gas Industry with Machine Learning-Based Ids. Available online: https://ssrn.com/abstract=4630706 (accessed on 26 September 2024).
26. Pettersen, S.; Grøtan, T.O. Exploring the grounds for cyber resilience in the hyper-connected oil and gas industry. *Saf. Sci.* **2024**, *171*, 106384. [CrossRef]
27. Obonna, U.O.; Opara, F.K.; Mbaocha, C.C.; Obichere, J.K.C.; Akwukwaegbu, I.O.; Amaefule, M.M.; Nwakanma, C.I. Detection of Man-in-the-Middle (MitM) Cyber-Attacks in Oil and Gas Process Control Networks Using Machine Learning Algorithms. *Future Internet* **2023**, *15*, 280. [CrossRef]
28. Shohoud, M. Study the Effectiveness of ISO 27001 to Mitigate the Cyber Security Threats in the Egyptian Downstream Oil and Gas Industry. *J. Inf. Secur.* **2023**, *14*, 152–180. [CrossRef]
29. ISO 27001 Standard. Available online: https://www.iso.org/standard/27001 (accessed on 26 September 2024).
30. Progoulakis, I.; Nikitakos, N.; Rohmeyer, P.; Bunin, B.; Dalaklis, D.; Karamperidis, S. Perspectives on cyber security for offshore oil and gas assets. *J. Mar. Sci. Eng.* **2021**, *9*, 112. [CrossRef]
31. Avanzini, G.B.; Spessa, A. Cybersecurity verification approach for the oil & gas industry. In Proceedings of the Offshore Mediterranean Conference and Exhibition, Ravenna, Italy, 27–29 March 2019. Paper number OMC-2019.
32. Mohammed, A.S.; Anthi, E.; Rana, O.; Saxena, N.; Burnap, P. Detection and mitigation of field flooding attacks on oil and gas critical infrastructure communication. *Comput. Secur.* **2023**, *124*, 103007. [CrossRef]
33. Gueye, T.; Iqbal, A.; Wang, Y.; Mushtaq, R.T.; Petra, M.I. Bridging the Cybersecurity Gap: A Comprehensive Analysis of Threats to Power Systems, Water Storage, and Gas Network Industrial Control and Automation Systems. *Electronics* **2024**, *13*, 837. [CrossRef]
34. Patel, S. Cybersecurity in Electric Distribution: The One Weak Link in an Interconnected Power Grid and the Threat It Poses. *Georg. Wash. J. Energy Environ. Law* **2023**, *14*, 138.
35. Naiho HN, N.; Layode, O.; Adeleke, G.S.; Udeh, E.O.; Labake, T.T. Addressing cybersecurity challenges in smart grid technologies: Implications for sustainable energy infrastructure. *Eng. Sci. Technol. J.* **2024**, *5*, 1995–2015. [CrossRef]
36. Jiang, Y.; Jeusfeld, M.A.; Ding, J.; Sandahl, E. Model-Based Cybersecurity Analysis: Extending Enterprise Modeling to Critical Infrastructure Cybersecurity. *Bus. Inf. Syst. Eng.* **2023**, *65*, 643–676. [CrossRef]
37. Kazancı, B.A. The Strategic Importance of Cyber Security in Electric Energy Policies. *Int. J. Energy Econ. Policy* **2024**, *14*, 599–605. [CrossRef]
38. Ratnam, E.L.; Baldwin, K.G.; Mancarella, P.; Howden, M.; Seebeck, L. Electricity system resilience in a world of increased climate change and cybersecurity risk. *Electr. J.* **2020**, *33*, 106833. [CrossRef]
39. Sun, C.C.; Cardenas DJ, S.; Hahn, A.; Liu, C.C. Intrusion detection for cybersecurity of smart meters. *IEEE Trans. Smart Grid* **2020**, *12*, 612–622. [CrossRef]
40. Shaaban, M.; Tariq, U.; Ismail, M.; Almadani, N.A.; Mokhtar, M. Data-driven detection of electricity theft cyberattacks in PV generation. *IEEE Syst. J.* **2021**, *16*, 3349–3359. [CrossRef]
41. Ibrahem, M.I.; Mahmoud, M.M.; Alsolami, F.; Alasmary, W.; Al-Ghamdi AS, A.M.; Shen, X. Electricity-theft detection for change-and-transmit advanced metering infrastructure. *IEEE Internet Things J.* **2022**, *9*, 25565–25580. [CrossRef]
42. Tolba, A.; Al-Makhadmeh, Z. A cybersecurity user authentication approach for securing smart grid communications. *Sustain. Energy Technol. Assess.* **2021**, *46*, 101284. [CrossRef]
43. Johnson, J.; Anderson, B.; Wright, B.; Quiroz, J.; Berg, T.; Graves, R.; Daley, J.; Phan, K.; Kunz, M.; Pratt, R.; et al. *Cybersecurity for Electric Vehicle Charging Infrastructure*; No. SAND2022-9315; Sandia National Lab. (SNL-NM): Albuquerque, NM, USA, 2022.
44. Bai, Y.; Sun, H.; Zhang, L.; Wu, H. Hybrid CNN–Transformer Network for Electricity Theft Detection in Smart Grids. *Sensors* **2023**, *23*, 8405. [CrossRef]
45. Musleh, A.S.; Ahmed, J.; Ahmed, N.; Xu, H.; Chen, G.; Kerr, S.; Jha, S. Experimental Cybersecurity Evaluation of Distributed Solar Inverters: Vulnerabilities and Impacts on the Australian Grid. *IEEE Trans. Smart Grid* **2024**, *15*, 5139–5150. [CrossRef]
46. Erkek, İ.; Irmak, E. Enhancing Cybersecurity of a Hydroelectric Power Plant Using Its Digital Twin Model. In Proceedings of the 2024 12th International Conference on Smart Grid (icSmartGrid), Setubal, Portugal, 27–29 May 2024; pp. 372–377.
47. Ismail, M.; Shaaban, M.F.; Naidu, M.; Serpedin, E. Deep learning detection of electricity theft cyber-attacks in renewable distributed generation. *IEEE Trans. Smart Grid* **2020**, *11*, 3428–3437. [CrossRef]

48. Takiddin, A.; Ismail, M.; Zafar, U.; Serpedin, E. Variational auto-encoder-based detection of electricity stealth cyber-attacks in AMI networks. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 1590–1594.
49. Takiddin, A.; Ismail, M.; Nabil, M.; Mahmoud, M.M.; Serpedin, E. Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings. *IEEE Syst. J.* **2020**, *15*, 4189–4198. [CrossRef]
50. Tang, D.; Fang, Y.P.; Zio, E. Vulnerability analysis of demand-response with renewable energy integration in smart grids to cyber-attacks and online detection methods. *Reliab. Eng. Syst. Saf.* **2023**, *235*, 109212. [CrossRef]
51. Heymann, F.; Henry, S.; Galus, M. Cybersecurity and resilience in the swiss electricity sector: Status and policy options. *Util. Policy* **2022**, *79*, 101432. [CrossRef]
52. Ge, P.; Chen, B.; Teng, F. Cyber-Resilient Self-Triggered Distributed Control of Networked Microgrids Against Multi-Layer DoS Attacks. *IEEE Trans. Smart Grid* **2023**, *14*, 3114–3124. [CrossRef]
53. Liu, M.; Teng, F.; Zhang, Z.; Ge, P.; Sun, M.; Deng, R.; Cheng, P.; Chen, J. Enhancing Cyber-Resiliency of DER-Based Smart Grid: A Survey. *IEEE Trans. Smart Grid* **2024**, *15*, 4998–5030. [CrossRef]
54. Nafees, M.N.; Saxena, N.; Cardenas, A.; Grijalva, S.; Burnap, P. Smart Grid Cyber-Physical Situational Awareness of Complex Operational Technology Attacks: A Review. *ACM Comput. Surv.* **2023**, *55*, 215. [CrossRef]
55. Zhang, Z.; Liu, M.; Sun, M.; Deng, R.; Cheng, P.; Niyato, D.; Chow, M.Y.; Chen, J. Vulnerability of Machine Learning Approaches Applied in IoT-Based Smart Grid: A Review. *IEEE Internet Things J.* **2024**, *11*, 18951–18975. [CrossRef]
56. Inayat, U.; Zia, M.F.; Mahmood, S.; Berghout, T.; Benbouzid, M. Cybersecurity Enhancement of Smart Grid: Attacks, Methods, and Prospects. *Electronics* **2022**, *11*, 3854. [CrossRef]
57. Ayodeji, A.; Mohamed, M.; Li, L.; Di Buono, A.; Pierce, I.; Ahmed, H. Cyber security in the nuclear industry: A closer look at digital control systems, networks and human factors. *Prog. Nucl. Energy* **2023**, *161*, 104738. [CrossRef]
58. Greiman, V. Nuclear Cyber Attacks: A Study of Sabotage and Regulation of Critical Infrastructure. In Proceedings of the International Conference on Cyber Warfare and Security, Towson, MD, USA, 9–10 March 2023; Volume 18, pp. 103–110.
59. Falowo, O.I.; Kropczynski, J.; Li, C. Protecting Critical Infrastructure: Strategies for Managing Cybersecurity Risks in Nuclear Fusion Facilities. In Proceedings of the 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Wuhan, China, 21–24 December 2023; pp. 1050–1061.
60. Jung, D.; Shin, J.; Lee, C.; Kwon, K.; Seo, J.T. Cyber security controls in nuclear power plant by technical assessment methodology. *IEEE Access* **2023**, *11*, 15229–15241. [CrossRef]
61. Yockey, P.; Erickson, A.; Spirito, C. Cyber threat assessment of machine learning driven autonomous control systems of nuclear power plants. *Prog. Nucl. Energy* **2023**, *166*, 104960. [CrossRef]
62. Morstyn, T.; Hredzak, B.; Agelidis, V.G. Control strategies for microgrids with distributed energy storage systems: An overview. *IEEE Trans. Smart Grid* **2016**, *9*, 3652–3666. [CrossRef]
63. Blaabjerg, F.; Teodorescu, R.; Liserre, M.; Timbus, A.V. Overview of control and grid synchronization for distributed power generation systems. *IEEE Trans. Ind. Electron.* **2006**, *53*, 1398–1409. [CrossRef]
64. Khan, S.; Madnick, S.; Moulton, A. *Cybersafety Analysis of Industrial Control System for Gas Turbines*; Cybersecurity Interdisciplinary Systems Laboratory (CISL): Cambridge, MA, USA, 2018.
65. Coakley, D.; Raftery, P.; Keane, M. A review of methods to match building energy simulation models to measured data. *Renew. Sustain. Energy Rev.* **2014**, *37*, 123–141. [CrossRef]
66. Bhattarai, B.P.; Paudyal, S.; Luo, Y.; Mohanpurkar, M.; Cheung, K.; Tonkoski, R.; Zhang, X. Big data analytics in smart grids: State-of-the-art, challenges, opportunities, and future directions. *IET Smart Grid* **2019**, *2*, 141–154. [CrossRef]
67. vom Scheidt, F.; Medinová, H.; Ludwig, N.; Richter, B.; Staudt, P.; Weinhardt, C. Data analytics in the electricity sector–A quantitative and qualitative literature review. *Energy AI* **2020**, *1*, 100009. [CrossRef]
68. Available online: https://www.ieee.org/ (accessed on 20 June 2024).
69. Available online: https://www.acm.org/ (accessed on 20 June 2024).
70. Fernandez, E.B.; Brazhuk, A. A critical analysis of Zero Trust Architecture (ZTA). *Comput. Stand. Interfaces* **2024**, *89*, 103832. [CrossRef]
71. Lombardi, F.; Fanton, A. From DevOps to DevSecOps is not enough. CyberDevOps: An extreme shifting-left architecture to bring cybersecurity within software security lifecycle pipeline. *Softw. Qual. J.* **2023**, *31*, 619–654. [CrossRef]
72. Nygård, A.R.; Katsikas, S.K. Ethical hardware reverse engineering for securing the digital supply chain in critical infrastructure. *Inf. Comput. Secur.* **2024**, *32*, 365–377. [CrossRef]
73. Al Obaidan, F.; Saeed, S. Digital transformation and cybersecurity challenges: A study of malware detection using machine learning techniques. In *Handbook of Research on Advancing Cybersecurity for Digital Transformation*; IGI Global: Hershey, PA, USA, 2021; pp. 203–226.
74. Alohali, M.A.; Alasmari, N.; Maashi, M.; Nouri, A.M.; Rizwanullah, M.; Yaseen, I.; Alneil, A.A. Metaheuristics with deep learning driven phishing detection for sustainable and secure environment. *Sustain. Energy Technol. Assess.* **2023**, *56*, 103114.
75. Georgiadou, A.; Michalitsi-Psarrou, A.; Askounis, D. A security awareness and competency evaluation in the energy sector. *Comput. Secur.* **2023**, *129*, 103199. [CrossRef]

76. Sharma, A.; Gupta, B.B.; Singh, A.K.; Saraswat, V.K. Advanced persistent threats (apt): Evolution, anatomy, attribution and countermeasures. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 9355–9381. [CrossRef]
77. Guo, Y. A review of Machine Learning-based zero-day attack detection: Challenges and future directions. *Comput. Commun.* **2023**, *198*, 175–185. [CrossRef] [PubMed]
78. Alzaabi, F.R.; Mehmood, A. A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods. *IEEE Access* **2024**, *12*, 30907–30927. [CrossRef]
79. Kurii, Y.; Opirskyy, I. Analysis and Comparison of the NIST SP 800-53 and ISO/IEC 27001: 2013. In Proceedings of the CPITS-2022: Cybersecurity Providing in Information and Telecommunication Systems, Kyiv, Ukraine, 13 October 2022.
80. Kitsios, F.; Chatzidimitriou, E.; Kamariotou, M. The ISO/IEC 27001 information security management standard: How to extract value from data in the IT sector. *Sustainability* **2023**, *15*, 5828. [CrossRef]
81. Heluany, J.B.; Galvão, R. IEC 62443 standard for hydro power plants. *Energies* **2023**, *16*, 1452. [CrossRef]
82. Mortensen, B.O.G.; Hjerrild, L. Legal Overview of Latest Developments in the Energy Sector Regarding Data Protection and Cybersecurity. In Proceedings of the Energy Informatics Academy Conference, Kuta, Bali, Indonesia, 23–25 October 2024; Springer Nature: Cham, Switzerland, 2025; pp. 112–119.
83. Avramidou, M.; Biasin, E.; Kamenjasevic, E.; Kun, E.; Nisevic, M. Cybersecurity and the NIS2 Directive: Regulatory aspects and sectoral perspectives. In Proceedings of the Second ECSCI Workshop on Critical Infrastructure Protection and Resilience, Online, 27–29 April 2022; Steinbeis-Edition: Stuttgart, Germany, 2023; pp. 91–92.
84. Available online: https://www.energy.gov/ceser/presidential-policy-directive-21 (accessed on 25 July 2024).
85. Khubrani, M.M.; Alam, S. Blockchain-based microgrid for safe and reliable power generation and distribution: A case study of Saudi Arabia. *Energies* **2023**, *16*, 5963. [CrossRef]
86. Mangla, C.; Rani, S.; Qureshi NM, F.; Singh, A. Mitigating 5G security challenges for next-gen industry using quantum computing. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101334. [CrossRef]
87. Siluk JC, M.; de Carvalho, P.S.; Thomasi, V.; Pappis CD, O.; Schaefer, J.L. Cloud-based energy management systems: Terminologies, concepts and definitions. *Energy Res. Soc. Sci.* **2023**, *106*, 103313. [CrossRef]
88. Carreiro, A.; Silva, C.; Antunes, M. The use of gamification on cybersecurity awareness of healthcare professionals. *Procedia Comput. Sci.* **2024**, *239*, 526–533. [CrossRef]
89. Saeed, S.; Suayyid, S.A.; Al-Ghamdi, M.S.; Al-Muhaisen, H.; Almuhaideb, A.M. A systematic literature review on cyber threat intelligence for organizational cybersecurity resilience. *Sensors* **2023**, *23*, 7273. [CrossRef] [PubMed]
90. Mohammed, A. Detection and Mitigation Strategies for Cyber-Attacks in Offshore Oil and Gas Industrial Networks. Ph.D. Dissertation, Cardiff University, Cardiff, UK, 2024.
91. Saeed, S. Digital Workplaces and Information Security Behavior of Business Employees: An Empirical Study of Saudi Arabia. *Sustainability* **2023**, *15*, 6019. [CrossRef]
92. Ahmad, A.; Maynard, S.B.; Motahhir, S.; Anderson, A. Case-based learning in the management practice of information security: An innovative pedagogical instrument. *Pers. Ubiquitous Comput.* **2021**, *25*, 853–877. [CrossRef]
93. Patterson, C.M.; Nurse, J.R.; Franqueira, V.N. "I don't think we're there yet": The practices and challenges of organisational learning from cyber security incidents. *Comput. Secur.* **2024**, *139*, 103699. [CrossRef]
94. Hussain, M. An Effective Cybersecurity Risk Assessment Framework for a Public Sector Gas Production/Distribution Company. Ph.D. Dissertation, National College of Ireland, Dublin, Ireland, 2023.
95. Bergset, S.; Nyland, A.J. Ensuring Safe and Secure Operations in the Norwegian Petroleum Industry: A Study on Assessing Trends in Cyber Risk Levels. Master's Thesis, NTNU, Trondheim, Norway, 2023.
96. Leppäsalo, N. Enhancing Cybersecurity Considerations in Plant-Level Safety Design of Nuclear Power Plant. Master's Thesis, Aalto University, Espoo, Finland, 2024.
97. Lee, S.W.; Lee, J.H. Improving the Efficiency of Cybersecurity Risk Analysis Methods for Nuclear Power Plant Control Systems. *J. Korea Inst. Inf. Secur. Cryptol.* **2024**, *34*, 537–552.
98. Fuller, C.R. Shortening the Skills Gap: An Exploratory Study of Cybersecurity Professional Experience. Ph.D. Thesis, Capella University, Minneapolis, MN, USA, 2016.
99. Anderson, A.; Ahmad, A.; Chang, S. Case-Based Learning for Cybersecurity Leaders: A Systematic Review and Research Agenda. *Inf. Manag.* **2024**, *61*, 104015. [CrossRef]
100. Saeed, S.; Bajwa, I.S.; Mahmood, Z. *Human Factors in Software Development and Design*; IGI Global: Hershey, PA, USA, 2014.

information

MDPI

Perspective

Improving the Cybersecurity Awareness of Young Adults through a Game-Based Informal Learning Strategy

Giorgia Tempestini, Sara Merà, Marco Pietro Palange, Alexandra Bucciarelli and Francesco Di Nocera *

Department of Planning, Design and Technology of Architecture, Sapienza University of Rome, 00196 Rome, Italy
* Correspondence: francesco.dinocera@uniroma1.it

Abstract: Knowing about a danger is not enough to avoid it. Our daily lives offer countless examples of occasions in which we act imprudently for various reasons, even though we know we are taking risks. Nevertheless, circumstances in which we lack the necessary knowledge can lead us to run into unpleasant or harmful situations without being aware of it. In cybersecurity, knowledge of the dangers (as well as the mechanics of a possible attack) makes a huge difference. This is why specific training is provided in organizations, along with awareness campaigns. However, security training is often generic, boring, and a mere fulfillment of obligations rather than a tool for behavioral change. Today, we can deliver content through various devices and platforms that people access for both work and leisure, so that learning can happen incidentally and with almost no effort. Distributing knowledge in small, dedicated units creates the conditions for lasting, effective learning and is more effective than teaching through traditional courses (whether delivered in-person or online). In this article, we present an ongoing project on cybersecurity informal learning, including the design of a small video game. The intervention is aimed at helping young adults (18–25 years) to understand the mechanics of cookies and their role in the dynamics of cyberattacks. Consistent with the idea that a comprehensive course may be unsuitable for delivering cybersecurity training, the game covers and deliberately limits itself to that topic only. We also provide detailed considerations related to the evaluation of its effectiveness, although this is outside the scope of the present paper.

Keywords: cybersecurity; behavior; knowledge; learning; games; gamification; cookies

1. Introduction

The human element is a critical weak point in the Information Technology (IT) security chain [1,2]. People are vulnerable to cybersecurity threats, either because they have limited knowledge or because, despite their awareness, they apply "only minimal protective measures, usually relatively common and simple ones" ([3], p.82), if any. Here, we are not concerned with the distinction between knowledge and awareness, which some authors consider to be separate constructs. For the purposes of this paper, we will consider knowledge to be indicative of awareness, whether or not it leads to appropriate behavior. Knowledge alone is not sufficient to explain behavior. For example, Lorenz and colleagues [4] observed that when information was provided on how to create passwords, only 2% followed the instructions at follow-up. Nevertheless, a lack of knowledge is usually associated with less attention being paid to cybersecurity threats. For example, Tempestini and colleagues [5] observed how those who have poorer knowledge of procedures, the consequences of their actions, policies, and their responsibilities are less careful about the actions they take, and do not protect their devices from different types of threats. Moreover, Di Nocera and colleagues [6] demonstrated that poor cybersecurity knowledge is associated with ignoring good practices, like using two-factor authentication.

While attitudes toward risk are not easy to change, knowledge can be easily improved. Indeed, in response to the exponential growth of security threats, a great deal of cybersecurity training is currently being offered. These courses target two different types of audience:

some courses are aimed at students or professionals in the field, while others are intended for non-technical employees to minimize the risk of insecure IT behavior. Both types have been addressed in the literature with the final objective of providing recommendations and guidelines to help design better cybersecurity courses. For example, González-Manzano and De Fuentes [7] analyzed 35 free courses, offered on various platforms (i.e., Coursera, Cybrary.it, edX, and Udacity) and aimed at individuals with various levels of knowledge (beginner, intermediate, and advanced). The authors provided indications relating to the contents, the delivery methods, the teachers' attitude towards the topics covered, the personalization of the courses, and the duration of the modules. In another work, Payne et al. [8] presented a detailed description of a cybersecurity course on which they based their recommendations for the design of future courses. Their contribution highlights, above all, the importance of the interdisciplinarity of the courses. This is an aspect that is certainly important, but which needs to be tested empirically.

Although it is interesting to analyze the literature devoted to professional training, our focus is on organizational cybersecurity training delivered along with (or as part of) awareness programs. In this area, the literature points to several limitations, particularly with respect to employees' perceptions of the courses themselves. In fact, employees often believe that the courses are boring and not very personalized to the needs of their organization, and there is a lack of incentive from the organization to the employees to participate in the course [9].

2. Background

Among the papers suggesting guidelines and best practices, few studies have empirically examined the effectiveness of cybersecurity training. He and colleagues [10] investigated the effect of different cybersecurity training methods on employees' cybersecurity risk perception and self-reported behavior. Employees were divided into four groups: a group who watched short educational videos, a group who read four reports, a group who watched videos and read reports, and a control group who received no intervention. All participants were required to complete a pre- and post-intervention test, specifically devised by the authors to assess perceived vulnerability, perceived severity, perceived benefits, perceived barriers, response efficacy, response cost, security self-efficacy, and behavioral intention. The study found positive effects for all methods with respect to vulnerability, barriers, response costs, self-efficacy, and behavioral intentions. Moreover, they found that evidence-based malware reporting is a relatively better training method than the other two training methods: people remember information more easily when it is presented in a way that makes it personally relevant.

Recently, Prummer and colleagues [11] investigated various learning modes in a review in which 142 articles (empirical articles, speculative articles, and intervention proposals) were analyzed. The authors identified the following types of training: game-based, presentation-based, simulation-based, information-based, video-based, text-based, and discussion-based. The authors compared all the courses by analyzing specific aspects, including the following:

- Properties: online or in person, group or individual;
- The theories on which the courses were based: Protection Motivation Theory, Theory of Planned Behavior, Theory of Reasoned Action, Signal Detection Theory, and General Deterrence Theory;
- The targets at which the courses were aimed: employees, students, young adults, or the general population;
- The effects of the training: fun, usefulness, or effectiveness.

The latter criterion is crucial because it is helpful for clarifying the actual utility of the different methods. Several studies among those analyzed in the review reported positive feedback, especially when techniques such as game-based or simulation-based training were used. Regarding effectiveness, the results showed an increase in almost all cases (except for in five articles), albeit with some differences depending on the type of training

used. Abawajy [12], for example, analyzed and compared the effects of game-, video-, and text-based training and found that awareness rates increased significantly across all conditions. It was also found that participants showed improvements in different areas depending on the training method they were assigned to. For example, game-based intervention led to an increase in the ability to identify scam sites, whereas video-based intervention increased knowledge about phishing.

2.1. Informal Learning

It appears clear that knowledge gained through formal training is not always sufficient to make users adopt secure behaviors. Formal training programs cannot always adequately prepare people to generalize their knowledge to all possible scenarios and are typically not designed to equip individuals for continuous learning.

Therefore, it becomes essential to explore alternative methods of delivering cybersecurity education that can be adapted to real-world scenarios and foster continuous and autonomous, self-paced learning. Currently, we can deliver content across various devices and platforms that people access for both work and leisure, so that learning about a great variety of topics can happen casually and almost effortlessly. Research suggests that a significant amount of learning happens through informal activities [13]. These activities can create an environment that fosters deeper and more effective learning compared to traditional courses. While traditional courses offer valuable information, they may not always be engaging for everyone. By incorporating informal learning methods, it would be possible to create a more well-rounded educational experience that promotes lasting behavioral change. Marsick and Watkins [14] actually make a distinction between "informal learning" and "incidental learning", distinguishing them on the basis of intentionality. People may learn informally, while intentionally choosing to seek out and learn new ideas, albeit in a less structured way than in formal settings. In contrast, incidental or implicit learning occurs during everyday activities, without a conscious attempt to learn or an awareness of what has been learned. However, this distinction is not made by everyone, and the two terms are often used interchangeably.

Here, with the expression "informal learning", we are referring to all those methods that take place in an unstructured way, often through daily experiences, social interactions, and environmental resources. It is characterized by the speed with which new knowledge can be acquired in response to emerging needs, and often does not entail significant costs. The importance of informal training has been documented in various contexts, both at the workplace [15] and in schools [16]. Many advantages have been found, including flexibility, the practical application of acquired skills, self-direction [14], and a greater level of involvement, especially with adolescents [17]. While there are numerous articles that deal with informal learning, its advantages, and its peculiarities, there are not many contributions specifically addressing the topic of informal learning in the field of information computer security. Rader and Wash [18] compared three different forms of informal learning related to computer security (news articles, web pages containing computer security advice, and stories about the experiences of friends and family). They observed that information provided by peers focuses mainly on the subjects who conduct the attacks, the information provided by web pages focuses instead on how the attacks are conducted, and the information coming from the news focuses on the consequences. Rader et al. [19] dealt with "stories" as informal lessons about security. Stories about others reveal useful information and it is easier to make our way through our complex world if we can learn from the experiences of others. Users reported that learning about other people's stories changed the way people think about security and act in similar occasions. The study was replicated 10 years later by Pfeffer et al. [20], who updated the contents to reflect contemporary technological advancements. While confirming the importance of storytelling as a means of informal learning, the authors also observed how storytelling worked when delivered via social media. They also observed how younger and more

educated participants perceive threats as less serious, probably because they have been more exposed to security news and therefore perceive security threats as less impactful.

Informal learning in a context such as cybersecurity is critical for several reasons: First, cybersecurity evolves constantly, with threats and vulnerabilities emerging rapidly and becoming more sophisticated, especially with the advancement of AI. Second, informal learning involves the application of knowledge to real life. For example, in competitions such as Capture the Flag (CTF), participants must—individually or as part of a team—test their knowledge by solving some security problems, which can range from exploiting websites to breaching unsecured networks [21]. The use of CTFs has been shown to be an effective way of improving cybersecurity education through gamification [22].

From our perspective, an informal learning approach is based on the idea that, rather than trying to categorize topics and merge many specific issues into formal training modules, cybersecurity education is more effective if information on each specific threat (e.g., phishing emails) or countermeasure (e.g., multiple factor authentication) is delivered in single "learning pills". Each "pill" may repeat different things to settle knowledge and may have a convenient outlet and modality to assist delivery. The proposal is very straightforward. If we need to inform people that do not know about phishing emails, we are obviously targeting people who do not know about this from other sources and have not looked it up themselves. We should expect them not to have even basic knowledge about cybersecurity. Formal training might be out of their comfort zone, being considered too complicated or too tedious. A communication campaign within the organization, in which examples of phishing emails are publicly dissected, may be much more informative than a formal training explaining the details of phishing (Figure 1a–c). The best outlet for such a campaign mostly depends on the type of target being addressed. Posters in the corridors of a school or a company as well as carousel posts on social media are examples of interventions that need less than a minute to be "consumed" by the user and that can be recalled more easily than a learning module on the concept of phishing. Humor could also be used to convey meaning and make the content highly shareable on social media. Other forms of content may have other more appropriate outlets. For example, some content may be easily shared using simple games when targeting people who normally spend time playing them. Here, the idea is to let people stumble upon content so that their knowledge changes from zero ("I didn't know that!") to where they have the ability to recognize the threat. What we are devising here is "molecular training", which is delivered informally, repeatedly, and through various media, contexts, and situations chosen from among those more appropriate for the type of user being targeted.

2.2. Serious Games

The literature on using game strategies for learning is highly diverse, with multiple research communities exploring this topic. A significant portion of this research falls under the umbrella of "serious gaming". To better understand the distinction between informal learning (described earlier) and serious gaming, it is useful to outline some key concepts that define this area. It is worth noting that many times "serious games" and "gamification" are used interchangeably. However, we prefer to use the term gamification to describe a process whose roots are in the functional analysis of behavior (see Section 2.3).

Serious gaming refers to games that are specifically designed with the primary goal of educating, training, or solving problems, rather than simply providing entertainment. These games employ game mechanics to engage users while imparting specific knowledge or skills [23]. While serious games have long been recognized as effective tools in fields such as education, healthcare, business, and military training, their potential in cybersecurity has only recently begun to be explored. Early research in this area mainly focused on testing the effectiveness of serious games in cybersecurity education. For example, Hendrix et al. [24] conducted a literature review of serious games available at the time, classifying them by their creators (academia or industry) and by type (e.g., 2D point-and-click turn-based scenarios, 3D virtual worlds, and corporate contingency planning). However, Hendrix

found that these games were primarily used for short-term education, which is generally less effective at driving long-term behavioral change.

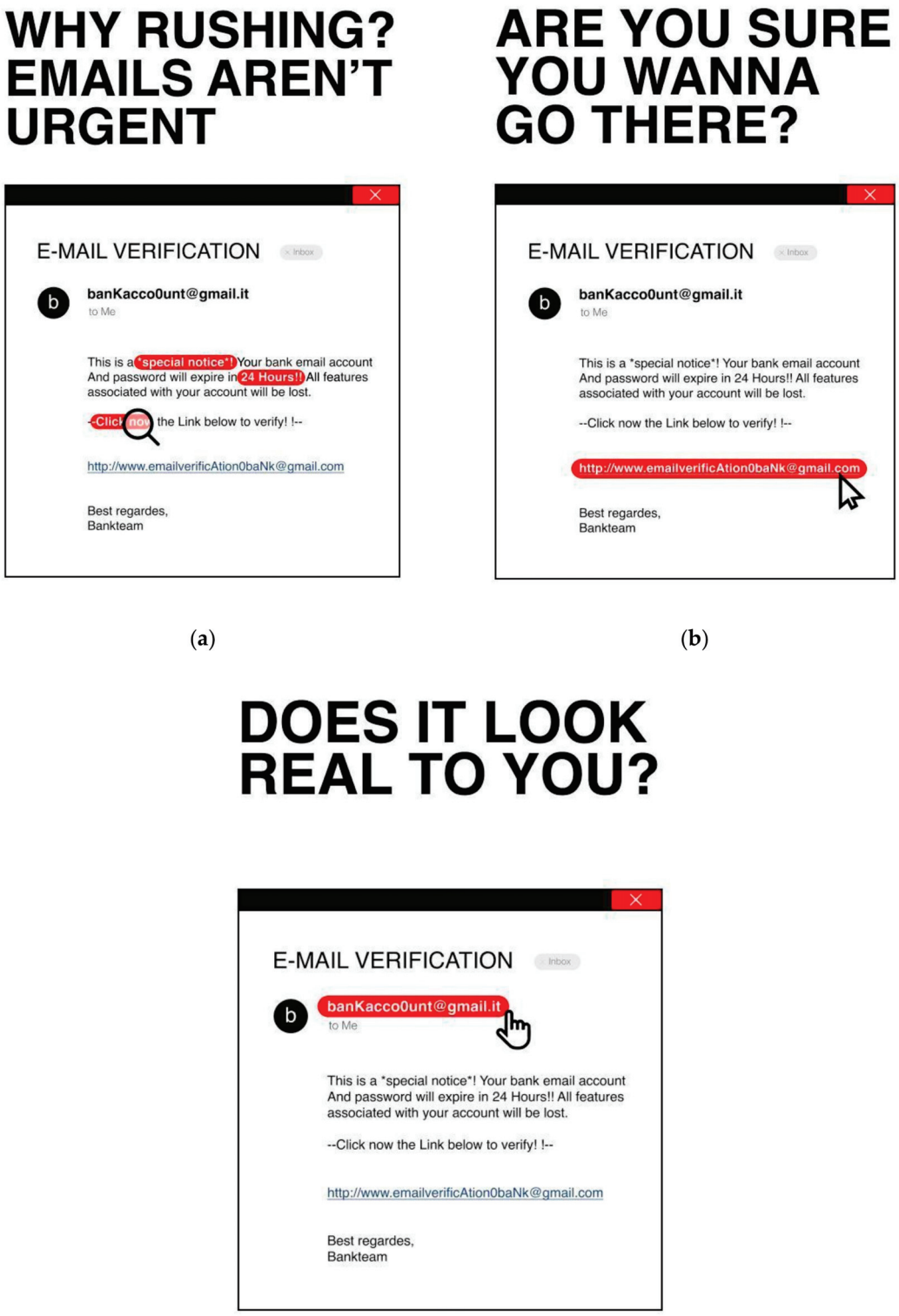

Figure 1. An example of a fictional communication campaign on the subject of phishing. There is emphasis on (**a**) a text that signals urgency, (**b**) a suspicious link within an email, and (**c**) a suspicious sender's email address.

Similarly, Alotaibi et al. [25] reviewed serious games in the field of cybersecurity and noted positive outcomes in terms of education and awareness. However, they also highlighted limitations, such as small sample sizes and content that was often too broad, making the games less effective for training in specific contexts. Hill et al. [26] echoed these concerns in their study, which aimed to provide an overview of existing cybersecurity games, particularly those focused on threats and protection, anti-phishing, and device security and privacy. They found that many of these games suffered from information overload and repetitive content, causing players to lose interest as they progressed through the game.

An analysis of recent studies on serious games reveals various interpretations of the concept. For example, Kulshrestha et al. [27] conducted a preliminary analysis of a serious game aimed at teaching network firewall concepts, using both objective and subjective methods. The objective analysis involved tracking player interactions, while the subjective analysis was based on user experience surveys. They also employed the learning mechanics–game mechanics framework to examine the relationship between game elements and learning outcomes. However, their study focused more on evaluating the quality of data using their hybrid approach than on assessing the game's overall effectiveness.

Much of the existing literature focuses on developing serious games for individuals with technical backgrounds in computer security. Coenraad et al. (2020) [28] noted that there are few games designed for users who do not consider themselves professionals in the field. One such study that addressed this gap is by Hart and colleagues [29], who introduced a board game called Risko designed to educate employees without technical training about the risks of cyberattacks and strategies for self-defense. The authors chose a card-based format, grounded in constructivist theory, and referenced existing games whose effectiveness had already been evaluated. In Risko, one player assumes the role of the attacker each turn, while the others propose countermeasures. Both attacks and defenses are inspired by industry and government standards, allowing the game to be adapted to various contexts and scenarios. A game master facilitates learning by encouraging players to reflect on their strategies, fostering discussion, and providing immediate feedback on the effectiveness of their decisions. The authors conducted a thorough analysis of existing board games, identifying potential flaws they should avoid in their own designs. Subsequent experiments confirmed the effectiveness of the game [29].

In contrast, Jaffray and colleagues [30] introduced SherLOCKED, a 2D, multi-level, top-down detective-themed game. Players control a detective, exploring rooms to find objects and answer related questions. The authors argue that role-playing games are often valuable and preferred by their target audience—students.

Other researchers have compared different types of serious games to determine which approaches are most effective. For example, Gaurav et al. [31] compared two games: one focused on teaching firewall protection against unauthorized access and another illustrating how SQL injection attacks compromise online databases. Both games were developed in two versions: a non-adaptive version that did not adjust to the player's abilities and a machine learning-based adaptive version. In the adaptive version, a machine learning agent classified players as beginners or experts based on their gaming behavior and adjusted the game's difficulty accordingly. The goal was to keep players consistently challenged without making the game too easy or too difficult. The results indicated that well-designed adaptive games can lead to better learning outcomes and a more satisfying user experience by catering to the specific needs of each player.

This brief overview reveals that the field of "serious gaming" is still evolving and lacks unified design frameworks and standardized methodologies for evaluating the effectiveness of interventions.

2.3. A Behavioral Account

So far, we have described two approaches to informal learning, one based on communication and the other on gaming. Both these approaches can be addressed in behavioral terms. Indeed, the use of verbal descriptions of behaviors and their consequences (positive or negative), as in the examples reported in Figure 1, represents an instance of learning based on rule-governed behaviors. In contrast, gaming represents a situation in which users are both rewarded for appropriate behavior (positive reinforcement) and encouraged to avoid conditions leading to a loss (negative reinforcement). This latter situation represents a form of learning based on contingencies. The distinction between rule-governed behavior and contingency-shaped behavior was first made by Skinner in the late 1960s [32,33]. He pointed out that, in the first form of learning, there is an instructional episode, the presentation of an instruction, a response provoked by the instruction, and a consequence provided by an instructional agent as a function of its fulfillment. Rules are stimuli that specify, either directly or indirectly, consequences for behavior. In contrast, contingency-shaped behavior is behavior directly controlled by the relations between responses and their consequences. Nevertheless, behavior may also come under the control of antecedent stimuli. These are stimuli whose presence causes responses to produce their consequences.

A comprehensive discussion of the principles underlying behavior analysis is out of the scope of this article. However, it is worth noting that the behavior analysis approach provides one of the most comprehensive accounts of learning. Behavioral principles provide a scientific understanding of how people and animals learn to deal with complex situations, without referring to theories or models based on aspects that are not fully observable [34]. Essentially, this perspective emphasizes the importance of there being consequences for behavior. In practice, a behavior is more likely to be repeated if it is "reinforced", that is, when it is followed by beneficial consequences for the individual. Beneficial consequences can involve either the acquisition of beneficial stimulation (positive reinforcement) or the removal of aversive stimulation for the individual (negative reinforcement).

Although the technicalities of behavior analysis are often not adhered to, these principles are actually already integrated into many software applications under the name of gamification. The term gamification began to grow in popularity as late as 2010, when Deterding and colleagues [35] defined it as a strategy of incorporating game-like elements into non-game contexts, highlighting its relevance as a tool with which to advance learning objectives and motivate behavior change. The basic concept of gamification, however, has been around for much longer. In the literature, this is also referred to by other terms (e.g., serious games, persuasive games, and alternate reality games). Gamification involves the redesign of everyday activities based on the methods employed in game design. This redesign often involves socially relevant behavioral changes and, for that reason, has not gone unnoticed by behavior analysts. Skinner [36] had already noted how video games constituted an excellent example of contingency programming, in that players interact with a system of contingencies in which their behavior is guaranteed to be reinforced. By contacting salient and immediate consequences, players are almost guaranteed to succeed. In a nutshell, game playing is about organizing contingencies. As reported by Di Nocera and Tempestini [37], this is usually implemented using the so-called token economy, "a rather complex reinforcement system based on the accumulation of objects, namely tokens, that can be eventually exchanged for goods, services, or privileges" (p. 251).

2.4. Cookies Anyone?

In this article, we discuss an attempt to promote informal learning regarding computer cookies, a topic that is often overlooked by security training. This is an ongoing project still in its early phases of development. However, it is an appropriate example of what we are advocating here: the informal and short delivery of information on a specific topic, allowing people to gather new knowledge they would not seek independently (due to a lack of interest or any other reason). It should be clear that the choice of the topic is incidental here, and the same rationale applies to any other topic.

Computer "cookies" are text files used by websites to store information on user behavior while browsing. Cookies are usually harmless, and people do not pay much attention to dialog boxes that require them to accept cookies. Users typically click the default option without much thought [38]. Cookie banners are designed to inform users about the data being collected and to enhance the user experience. However, they are often perceived negatively due to their invasiveness and potential privacy violations. Their design frequently encourages users to accept cookies [39], as their omnipresence tends to annoy users [40]. This issue extends beyond just cookies. Due to habituation—the diminishing response to a frequently repeated stimulus—users often overlook crucial messages when repeatedly encountering security-related dialog boxes. Their response becomes automatic, leading them to click buttons merely to acknowledge the message without fully engaging with its content [41,42].

Cookies are also highly valuable pieces of information for malicious users. To illustrate their importance, a study by NordVPN [43] revealed that hackers have stolen 54 billion cookies, later distributing them on the Dark Web. Cookies involved in authentication procedures are particularly valuable, as these are used by browsers to identify users and verify their login status.

Recently, a critical exploit was discovered that allows the generation of persistent Google cookies through token manipulation. This exploit enables continuous access to Google services, even after a password reset. Another well-documented (https://github.com/mrd0x/WebView2-Cookie-Stealer, accessed on 1 July 2024) example comes from Mr.D0x, an anonymous penetration tester and security researcher who demonstrated how Microsoft Edge WebView2 can be exploited to steal authentication cookies, potentially bypassing multi-factor authentication when accessing stolen accounts.

These examples clearly show that cookies, typically considered harmless, can pose significant security threats. Increasing user awareness about the potential risks associated with cookies could encourage more cautious behavior regarding their use and management.

2.5. Target Population

In an increasingly fast-paced and digitally driven society, online security has become a key concern for all age groups. However, not all people are equally aware of or protected against cyber risks. In particular, young adults are a particularly vulnerable group due to a number of specific factors that make them more susceptible to cyberattacks than other segments of the population.

Young adulthood or late adolescence is a developmental phase spanning the ages of 18 to 25 [44]. During this transitional period, young people move from childhood and adolescent systems into adult-centered systems. The number of young adults spending their free time on gaming consoles, computers, and smartphones is increasing rapidly. As a result, they are the most likely group to encounter various cyber threats and other risks associated with Internet use. A study by Zhao et al. [45] showed that young adults have a good awareness of the risks associated with sharing inappropriate content and approaching people unknown to them, but demonstrated that they lack awareness of the risks associated with sharing personal information online. As reported by Alanazi et al. [46], young adults should be made aware of the risks posed by cyber threats that arise from neglecting online security practices. They need to understand how adopting proper cybersecurity behaviors can lead to positive outcomes, such as staying safe online. Cybersecurity education should include practical demonstrations of the essential security measures that should be practiced in everyday Internet use. Therefore, integrating cybersecurity education with activities that promote incidental learning is especially advantageous for this age group.

3. Game Design

As mentioned earlier, game-based training is gaining popularity as a method for raising awareness about cyber risks [11]. Several notable examples of educational games exist. For instance, Space Shelter (https://spaceshelter.withgoogle.com/intl/it_it/, ac-

cessed on 1 July 2024) is a web-based video game developed by Google to raise awareness about online privacy and security. The game educates users about safe web browsing and personal data protection through a fun and interactive experience. Players navigate a spaceship, choosing an astronaut avatar and advancing by correctly answering quizzes and completing mini-games on online security. Key game mechanics include quizzes, guided completion, and drag-and-drop activities. The game has achieved a high level of engagement with 450,000 unique users, 40% of whom completed the game in an average of 10 min, and is available in seven languages.

Another game aimed at young people is Datak (https://www.datak.ch/#/start, accessed on 1 July 2024). In this single-player game, the player is hired as an intern by the mayor of a city and tasked with managing the social media network, facing various daily dilemmas and time constraints, with advice from YouTube videos. The game addresses topics such as the role of the Internet in everyday life, social networks, user actions, state surveillance, and commerce.

A more ambitious game is Interland (https://beinternetawesome.withgoogle.com/es_es/interland, accessed on 1 July 2024), which is designed to help children (ages 6 to 13) learn key lessons about web safety through four different experiences: the River of Reality (distinguishing truth from falsehood), the Treasure Tower (guarding personal information), the Courteous Kingdom (spreading kindness), and the Responsible Mountain (using technology wisely). The game teaches users to recognize scams, understand phishing and how to report it, safeguard personal information, and create secure and memorable passwords.

All the games mentioned above share a variety of content aimed at providing comprehensive training. This contrasts with the approach we are emphasizing here, the "molecular" and effective transfer of knowledge on a very circumscribed topic.

The game described below targets young adults who frequently use the Internet and are likely to encounter various cyber threats and other risks associated with Internet use [46]. This audience comprises digital natives with a direct and personal relationship with technology and a preference for gamified, active, and competitive learning methods. Involving them in initiatives to raise awareness can promote responsible online behavior and provide them with essential knowledge to help manage their digital privacy effectively. The game is designed to be played on smartphones, which are chosen for their accessibility, familiarity, interactivity, portability, and connectivity.

Razali and colleagues [47] recently tested alongside experts a guideline for designing educational games (in their case, in the field of climate change). The guideline is based on 13 game elements, and it is general enough to be applied to other topics. Below, we will employ their game elements to describe the characteristics of Cookie Aware.

3.1. Goal

The game must have a clear objective so that players can easily understand how to win the game. These goals can be ascribed to what Starks [48] refers to as "in-game" goals and are distinct from the "real-life outcomes", which fall under what Razali and colleagues refer to as "Scope" (see Section 3.4). Therefore, the goal is to obtain citizenship in CookieLand by demonstrating knowledge about cookies. The achievement of this goal is signaled by earning badges, which visibly represent the completion of levels or objectives [49].

3.2. Narrative Content

Storytelling in games enables players to identify with the main character, fostering a stronger connection and increased engagement. Isbister and Schaffer [50] have noted that games with a rich narrative component can significantly enhance the user experience by promoting deeper immersion and involvement. In Cookie Aware, village exploration is the key element of storytelling. To gain citizenship in CookieLand, the main character (Bibi) must earn badges that certify their knowledge of cookies. This knowledge is acquired through interactions with non-playable characters (NPCs), the inhabitants of CookieLand, who share stories about their cookie-related experiences. The believability of these stories

can greatly influence player engagement, highlighting the crucial role of a well-crafted narrative in game design [51].

The story takes place in CookieLand (Figure 2), a peaceful village inhabited by various types of cookies. CookieLand exists within everyone's computers, where each cookie has its own unique personality and interests. The game's narrative follows Bibi, an explorer who wishes to move to CookieLand. However, Bibi's transfer is initially halted by the village mayor. According to the village rules, in order to gain citizenship, Bibi must first demonstrate knowledge about the inhabitants and the community's hierarchical relationships.

Figure 2. Graphic visualization of the game setting, CookieLand village.

Bibi begins the journey by speaking with the villagers, who share information and interesting facts about the four quiz challenges that must be completed. After successfully completing each challenge, Bibi earns a badge (Figures 3–5), an official document certifying knowledge advancement. Once Bibi has earned all four badges, the player can finally move to CookieLand.

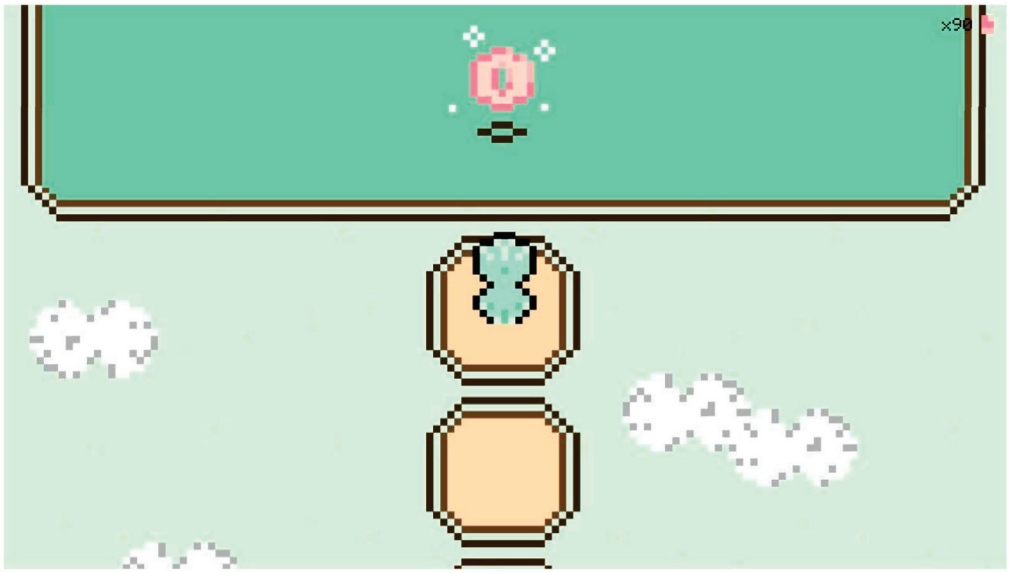

Figure 3. The moment at which the badge is obtained.

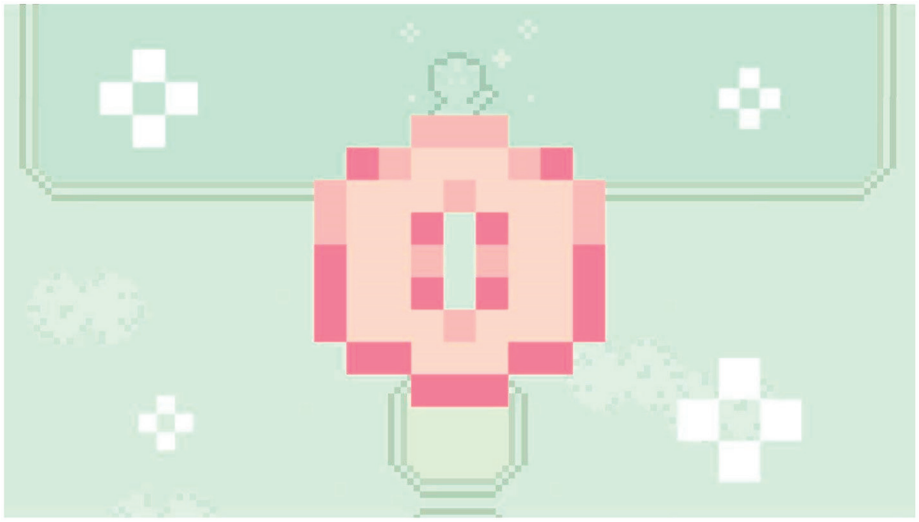

Figure 4. Badge graphics.

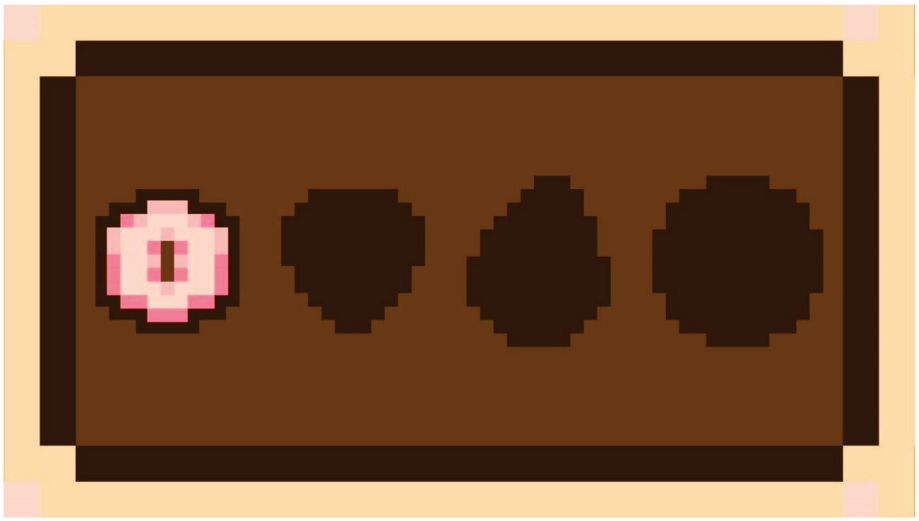

Figure 5. Badge insertion in the medallion box.

3.3. Rules

The game is structured into four levels, each requiring approximately 10 min to complete. To progress to the next level, the player must correctly answer questions related to the current level. Upon successfully completing a level, the player earns a badge. Collecting all the badges demonstrates the player's mastery and grants access to rewards, such as free selected products from vending machines.

3.4. Scope

The scope of the game is to enhance individuals' awareness and knowledge of cybersecurity. Specifically, the game aims to educate players on the importance of cookie management, explaining how cookies impact online privacy and security, and promoting responsible practices to protect personal data.

3.5. Genre

Cookie Aware blends two game genres to create an engaging experience. The first genre is the quiz, where players answer questions correctly to progress. Quizzes are known for their benefits in terms of learning and memory retention. Research by Butler and Roediger [52] highlighted how tests and quizzes can enhance information retention over time, demonstrating that testing is a powerful educational tool. In 2011, Butler and

Roediger further emphasized that the act of retrieving information during quizzes can improve long-term knowledge storage. Similarly, Blunt [53] explored how educational games, including quizzes, positively impact learning by boosting both engagement and memorization. There is a strong consensus that quizzes are effective learning tools.

The second genre featured in Cookie Aware is the role-playing game (RPG). RPGs enhance user immersion by encouraging players to identify with the game's protagonist, allowing for a deeper, more engaging experience. Role-playing games can enhance problem-solving skills by teaching players to first gather information and devise a strategy before attempting to solve a problem. In a longitudinal study, Adachi and Willoughby [54] showed that adolescents who played strategic video games across many years of high school reported steeper increases in self-reported problem-solving skills over time compared to participants who reported less sustained play.

3.6. Esthetic

The visual elements of the game are crucial in creating an engaging and memorable experience. Cookie Aware achieves this through a combination of pixel art graphics, pastel colors, and playful electronic sounds. The game's design draws inspiration from some of the most beloved classics in gaming, including *Pokémon Sapphire*, *The Legend of Zelda*, *Dragon Quest III*, and *Space Invaders*. Key visual elements from these games have been adapted and integrated to enhance the overall gaming experience.

Pixel art graphics were chosen for their simplicity and clarity, making the information easy to comprehend. Each character, icon, and scene is meticulously crafted with detailed pixel art, blending both playful and informative aspects. The typography features the Minecraft font, which is widely recognized due to its popularity.

Saturated, high-contrast colors were selected to ensure that there was a clear distinction between game elements, improving usability. The pastel palette creates a relaxing atmosphere, with blue and pink hues often used in confectionery advertising to offer a pleasant visual experience (Figure 6). Animations and transitions add dynamism to the gameplay. The initial vertical scrolling transition provides a panoramic view of the village, followed by a zoom-in on the protagonist entering the settlement. The game's instructions and objectives are then presented. Subsequent transitions track the protagonist's movements, zooming in on character interactions and changes in setting.

Figure 6. Color palette.

Playful electronic sounds further enhance the game's immersion. The soundtrack deepens the context of on-screen actions, while sound effects make the virtual world feel more realistic and responsive, providing immediate feedback to the player's actions and changes within the game environment.

3.7. Character Design

The game character represents the player within the game and often determines its success. In any game, the character plays a unique and crucial role, serving as the player's avatar as they navigate a virtual world filled with adventures and challenges [55]. However, a character holds little significance if it is not designed with a clear goal in mind. A game's storyline, conflicts, challenges, and atmosphere come together meaningfully when the main character has a specific objective for players to achieve at each level [56].

For the development of character design, Burgerman's [57] "20 Top Character Design Tips" served as a reference. The protagonist of Cookie Aware is Bibi, who resides within our electronic device but dreams of living in the beautiful town of CookieLand. Bibi is gender-neutral and designed without any specific characteristics that might hinder players' ability to empathize with them. Bibi's signature color is a light blue/teal, while the non-playable characters (NPCs) in CookieLand feature colors reminiscent of real cookies, making them easily recognizable and allowing Bibi to stand out within the setting.

3.8. Game Mode

Cookie Aware is designed as a single-player offline game where players earn badges by demonstrating their knowledge about cookies, enabling Bibi to eventually move into the village. The game features a scoring system, which ranks players and introduces an (online) element of competition. This can be especially beneficial at the class or school level, as it encourages greater participation in the game.

3.9. Game Level Design

The game's allocentric representation is based on a map constructed over a geometric grid. Only a portion of the world is displayed on the screen at any given time, encouraging players to explore further. This type of design is common in video game sagas such as Pokémon, Zelda, and Dragon Quest. The game features five distinct areas, each designed to introduce different concepts and information to the player.

3.10. Quiz

The game's narrative structure guides players through a series of quiz levels, where they receive immediate feedback and can earn badges. Forty questions were initially developed and divided across four levels to help users gain a solid understanding of cookies. Players start at the first level and advance to the next after successfully answering all questions in the current level. This progressive structure keeps players motivated and engaged, continuously challenging them to improve their performance.

The questions are varied, ranging from informative to ironic and friendly, making the experience both educational and entertaining. The initial questions are relatively simple, helping participants ease into the topic and build confidence. As the game progresses, the questions become more complex, requiring a deeper understanding, critical analysis, and practical application of the information. Each question offers four options, with only one correct answer (see the question example in Figure 7). Regardless of whether the player answers correctly, a detailed explanation is provided to deepen their understanding and prevent misconceptions (see Figure 8).

Each level focuses on a specific aspect of cookies, providing a comprehensive overview:

- Knowledge (Introduction to Cookies): The first level introduces players to the basics of cookies, covering key concepts such as what cookies are, where they come from, their primary purpose, and the types of information they store. This level also explains relevant cookie regulations, including the GDPR 2018 regulation. It serves as an essential foundation for players, helping them understand cookies before moving on to more complex topics in later levels.
- Types (Types of Cookies): In the second level, players learn to identify different types of cookies and understand the kinds of information they store. This level distinguishes between technical, profiling, and third-party cookies, as well as between session and

persistent cookies. It also explains the differences between hybrid and zombie cookies, highlighting how they differ from traditional cookies and the common issues caused by zombie cookies during browsing.

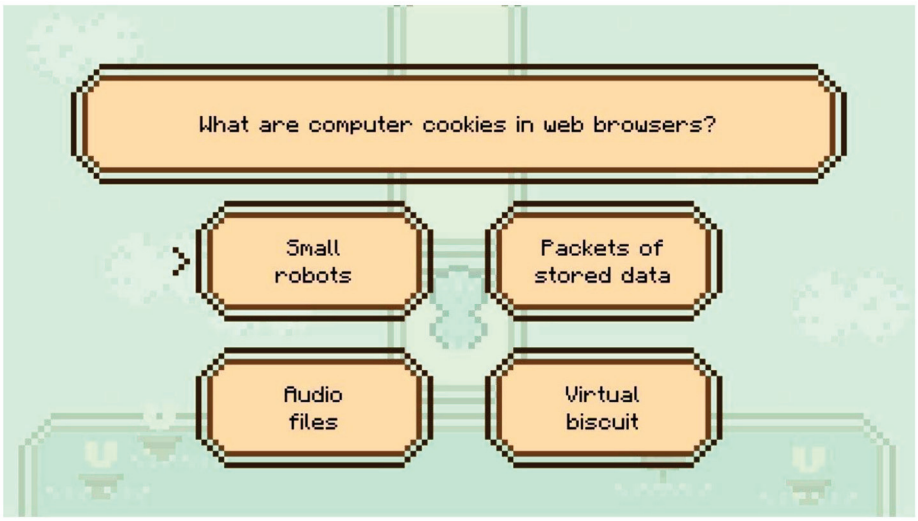

Figure 7. Example question with answer alternatives.

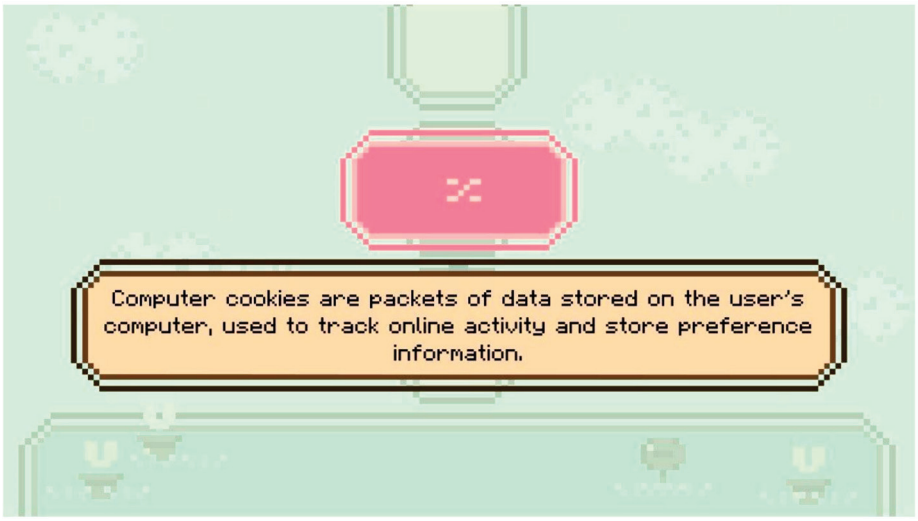

Figure 8. A more detailed explanation of the correct answer. Feedback is given both in the case of an incorrect answer and a correct answer by the player.

- Risks (Risks and Threats): The third level focuses on the potential risks associated with cookies, such as user tracking. It emphasizes the importance of being aware of these risks, particularly since cookies can handle sensitive information, including personal and financial data. This level also covers cyberattacks that exploit cookies, such as cookie poisoning and cross-site scripting (XSS), as well as the consequences of session cookie breaches and the best practices for safely managing cookies.
- Solutions (Solutions and Best Practices): The final level provides practical tips and strategies for safely managing cookies. It covers the importance of deleting cookies and offers guidance on how to control and manage them in browsers. This level also explains the functions of common cookie management tools, the effects of browsing in private or incognito mode, and explores safer alternatives to cookies. Additionally, it highlights the benefits of using the "I Don't Care About Cookies" extension and discusses how to identify GDPR-compliant sites, stressing the importance of reading privacy policies and cookie-related documents.

After correctly answering all 10 questions in a level, the level is considered complete (see Figure 9).

Figure 9. Screen upon the completion of the level.

3.11. Reward

Cookie Aware involves a token economy, an exchange system that provides immediate feedback on the appropriateness of individuals' behavior [58]. Tokens, or "exchangeables," are earned by achieving specific behavioral goals over a set period and can be traded for tangible rewards or privileges. Although tokens have no intrinsic value, they function as reinforcers—stimuli or events that increase the likelihood of a behavior when given in response to that behavior. If a stimulus does not result in a higher frequency of the target behavior, it is not considered a reinforcer. A key component of the token economy is the exchange of tokens for backup rewards.

The game will also feature a response cost system, taking away tokens when inappropriate behavior occurs. That will also add dynamism to the game.

3.12. Challenge

The game's challenge is driven by two mechanisms: (1) the progressively increasing difficulty of the questions within each level and during transitions between levels; (2) occasional pop-ups prompting choices to give consent to a great variety of permissions. Player responses will determine the dynamics of the game by allowing either a smooth exploration or disadvantages (e.g., obstacles, opponents).

3.13. Rank

Rank is an integral part of the game dynamics. The character's status changes from level to level, also changing some features of the character's appearance (e.g., color) and an update of the progress bar. As mentioned in Section 3.8, introducing a ranking system can encourage participation, with competition serving as a significant motivator [59]. The score obtained can also provide valuable feedback, helping players decide whether to replay a level to improve their ranking.

4. Discussion and Conclusions

The goal of this article was twofold. On the one hand, we wanted to argue for the preference of information/education "pills" over the more comprehensive courses usually offered in organizations to improve the awareness of non-expert users. On the other hand, we wanted to provide an example of a forthcoming intervention on a circumscribed topic that would take advantage of an informal/incidental learning mode.

From the literature review, two aspects seem to be salient in the use of informal learning strategies. The first is the need to define the audience for training. For example, adult employees of an organization and college students represent two different populations with different habits and approaches to information consumption. Second, it is necessary reduce the complexity of information/training units to ensure the learning and generalization of key concepts. Although any type of content can be delivered using informal learning approaches, the temptation to explain a phenomenon (e.g., phishing) in exhaustive detail must be resisted in favor of exposing the user to instances of a phenomenon (e.g., using mouseover to verify the nature of a link before clicking) that can then be more easily generalized to other situations. Traditional courses, by their nature, tend to consist of instructional modules that represent a more or less homogeneous organization of content. However, any categorization is reductive to the myriad of issues within the broad context of cybersecurity, some of which are often overlooked, even though they can have devastating consequences for the victims.

The concept is not new, and we can find similar approaches under the name of "microlearning" [60,61], in which complex information is broken down into smaller modules that are easy for the learner to use. However, microlearning still involves short lessons that can be accessed via mobile phone apps or via computers, whereas what we are advocating here is to create situations in which the users stumble upon content and increase their knowledge while they are doing something else. Social media offer an ideal venue for providing short, circumstantial content that is easy to consume, and that can be propagated through the mechanism of sharing. It is definitely an approach that should be considered in order to create awareness in a population and instill proper behavior. Indeed, this approach also qualifies as a strategy for creating verbal antecedents for rule-governed behavior. As an alternative to rule-governed behaviors, learning often occurs as a function of reinforcement contingencies. People enact behaviors that are followed by consequences, and these consequences reinforce that behavior.

Game environments are highly effective for establishing contingencies. The feedback received during gameplay, along with badges and accumulated points, provides positive reinforcement that helps to solidify learning. The use of a token economy is well documented in the literature as a strategy to keep the individual engaged in the learning process. Here, we have designed a simple game that can appeal to students and let them learn about a very specific topic usually overlooked by traditional cybersecurity training. Points and badges gained throughout the game experience could be used to gain status in classroom or school rankings, stimulating students to improve their performance (i.e., learning) while they are performing a leisure activity.

Future Directions

In this article, we presented a minimalist game to informally educate about cookies. Here, we explained its background, purpose, target audience, structure, content, game dynamics, and style. This is the first step in a project with many more phases. The next phases will focus on implementing the game and evaluating its effectiveness. The evaluation of strategies to improve cybersecurity awareness—and their potential impact on the adoption of appropriate behavior—should be a focus of research in this area. Cookie Aware will serve as a test bed for this, comparing the knowledge gained through this gaming platform with traditional teaching methods on the same topic.

A forthcoming study will assign participants to one of three experimental conditions: watching a video lesson, interacting directly with the game, and watching a video of a third person playing the game. The third condition is designed to assess the actual impact of the game experience beyond the content delivery method. All participants will complete the Cybersecurity Awareness Inventory [5], which will be used as a covariate to check prior cybersecurity knowledge. This study should provide valuable data on students' cybersecurity knowledge and establish a benchmark for future applications.

In addition, partnerships could be established with companies that operate vending machines in schools and colleges, as these machines often integrate smartphone applications that facilitate purchases. Cookie Aware could be seamlessly integrated into these applications, creating an innovative incentive for students to participate in the game by converting virtual rewards (points, badges) into tangible benefits such as free or discounted snacks and drinks from vending machines.

This partnership model could be customized for different institutions, allowing schools and universities to tailor rewards to their students' preferences, which could further increase engagement.

For educational institutions, this approach offers an innovative way to integrate gamified learning into daily routines, making cybersecurity education more relevant, interactive, and accessible. This collaboration could eventually serve as a model for other educational games that seek to combine learning with real-world incentives.

Cookie Aware could also be incorporated into high school projects focused on cybersecurity education. Such projects, aimed at teenagers, aim to develop the skills required for the safe and conscious use of the Internet and provide students with the resources to manage digital risks. These educational programs typically cover topics such as protecting personal information, managing strong passwords, recognizing online threats (such as phishing), and using social media responsibly. Cookie Aware, with its focus on educating users about cookies and privacy, aligns perfectly with these goals by providing tools that demonstrate how everyday online behavior can affect personal security.

In addition, Cookie Aware could facilitate discussions about digital ethics and privacy by encouraging students to think critically about the information they share online and the consequences of their actions. By incorporating Cookie Aware into cybersecurity education, schools can provide students with better tools to make informed decisions about their online presence, fostering a generation that is not only tech-savvy, but also aware of the importance of maintaining privacy and security in an increasingly connected world.

Author Contributions: The activity reported in this paper originates from a project work carried out by S.M., M.P.P. and A.B., under the supervision of F.D.N. and G.T., as part of their participation in class activities at the Master Program in Design, Multimedia, and Visual Communication, Sapienza University of Rome, Italy. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rahman, T.; Rohan, R.; Pal, D.; Kanthamanon, P. Human factors in cybersecurity: A scoping review. In Proceedings of the 12th International Conference on Advances in Information Technology, Bangkok, Thailand, 29 June–1 July 2021; pp. 1–11.
2. Alsharif, M.; Mishra, S.; AlShehri, M. Impact of Human Vulnerabilities on Cybersecurity. *Comput. Syst. Sci. Eng.* **2022**, *40*. [CrossRef]
3. Zwilling, M.; Klien, G.; Lesjak, D.; Wiechetek, Ł.; Cetin, F.; Basim, H.N. Cyber security awareness, knowledge and behavior: A comparative study. *J. Comput. Inf. Syst.* **2022**, *62*, 82–97. [CrossRef]
4. Lorenz, B.; Kikkas, K.; Klooster, A. The four most-used passwords are love, sex, secret, and god: Password security and training in different user groups. In Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust, Las Vegas, NV, USA, 21–26 July 2013; pp. 276–283.
5. Tempestini, G.; Rovira, E.; Pyke, A.; Di Nocera, F. The Cybersecurity Awareness INventory (CAIN): Early Phases of Development of a Tool for Assessing Cybersecurity Knowledge Based on the ISO/IEC 27032. *J. Cybersecur. Priv.* **2023**, *3*, 61–75. [CrossRef]
6. Di Nocera, F.; Tempestini, G.; Presaghi, F. Reliability and validity of the Cybersecurity Awareness INventory (CAIN). *Behav. Inf. Technol.* **2024**, 1–12. [CrossRef]

7. González-Manzano, L.; de Fuentes, J.M. Design recommendations for online cybersecurity courses. *Comput. Secur.* **2019**, *80*, 238–256. [CrossRef]
8. Payne, B.K.; He, W.; Wang, C.; Wittkower, D.E.; Wu, H. Cybersecurity, technology, and society: Developing an interdisciplinary, open, general education cybersecurity course. *J. Inf. Syst. Educ.* **2021**, *32*, 1334. [CrossRef]
9. He, W.; Zhang, Z. Enterprise cybersecurity training and awareness programs: Recommendations for success. *J. Organ. Comput. Electron. Commer.* **2019**, *29*, 249–257. [CrossRef]
10. He, W.; Ash, I.; Anwar, M.; Li, L.; Yuan, X.; Xu, L.; Tian, X. Improving employees' intellectual capacity for cybersecurity through evidence-based malware training. *J. Intellect. Cap.* **2020**, *21*, 203–213. [CrossRef]
11. Pruemmer, J.; van Steen, T.; van den Berg, B. A systematic review of current cybersecurity training methods. *Comput. Secur.* **2023**, *136*, 103585. [CrossRef]
12. Abawajy, J. User preference of cyber security awareness delivery methods. *Behav. Inf. Technol.* **2014**, *33*, 237–248. [CrossRef]
13. Cerasoli, C.P.; Alliger, G.M.; Donsbach, J.S.; Mathieu, J.E.; Tannenbaum, S.I.; Orvis, K.A. Antecedents and outcomes of informal learning behaviors: A meta-analysis. *J. Bus. Psychol.* **2018**, *33*, 203–230. [CrossRef]
14. Marsick, V.J.; Watkins, K. *Informal and Incidental Learning in the Workplace*; Routledge: New York, NY, USA, 2015.
15. Blume, B.D.; Ford, J.K.; Baldwin, T.T.; Huang, J.L. Transfer of training: A meta-analytic review. *J. Manag.* **2010**, *36*, 1065–1105. [CrossRef]
16. Lecat, A.; Raemdonck, I.; Beausaert, S.; März, V. The what and why of primary and secondary school teachers' informal learning activities. *Int. J. Educ. Res.* **2019**, *96*, 100–110. [CrossRef]
17. Mahoney, J.L.; Larson, R.W.; Eccles, J.S. (Eds.) *Organ. Act. as Context. Dev. Extracurricular Act. after-School Community Programs*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2005; pp. 3–22.
18. Rader, E.; Wash, R. Identifying patterns in informal sources of security information. *J. Cybersecur.* **2015**, *1*, 121–144. [CrossRef]
19. Rader, E.; Wash, R.; Brooks, B. Stories as informal lessons about security. In Proceedings of the Eighth Symposium on Usable Privacy and Security, Washington, DC, USA, 11–13 July 2012; pp. 1–17.
20. Pfeffer, K.; Mai, A.; Weippl, E.; Rader, E.; Krombholz, K. Replication: Stories as informal lessons about security. In Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022), Boston, MA, USA, 7–9 August 2022; pp. 1–18.
21. Švábenský, V.; Čeleda, P.; Vykopal, J.; Brišáková, S. Cybersecurity knowledge and skills taught in capture the flag challenges. *Comput. Secur.* **2021**, *102*, 102154. [CrossRef]
22. Balon, T.; Baggili, I. Cyber Competitions: A survey of competitions, tools, and systems to support cybersecurity education. *Educ. Inf. Technol.* **2023**, *28*, 11759–11791. [CrossRef]
23. Breuer, J.; Bente, G. Why so serious? On the relation of serious games and learning. *J. Comput. Game Cult.* **2010**, *4*, 7–24. [CrossRef]
24. Hendrix, M.; Al-Sherbaz, A.; Victoria, B. Game based cyber security training: Are serious games suitable for cyber security training? *Int. J. Serious Games* **2016**, *3*, 53–61. [CrossRef]
25. Alotaibi, F.; Furnell, S.; Stengel, I.; Papadaki, M. A review of using gaming technology for cyber-security awareness. *Int. J. Inf. Secur. Res. (IJISR)* **2016**, *6*, 660–666. [CrossRef]
26. Hill, W.A., Jr.; Fanuel, M.; Yuan, X.; Zhang, J.; Sajad, S. A survey of serious games for cybersecurity education and training. *KSU Proc. Cybersecur. Educ. Res. Pract.* **2020**, *7*, 1–17.
27. Kulshrestha, S.; Agrawal, S.; Gaurav, D.; Chaturvedi, M.; Sharma, S.; Bose, R. Development and validation of serious games for teaching cybersecurity. In Proceedings of the Serious Games: Joint International Conference, JCSG 2021, Virtual Event, 12–13 January 2022; Proceedings 7. Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 247–262.
28. Coenraad, M.; Pellicone, A.; Ketelhut, D.J.; Cukier, M.; Plane, J.; Weintrop, D. Experiencing cybersecurity one game at a time: A systematic review of cybersecurity digital games. *Simul. Gaming* **2020**, *51*, 586–611. [CrossRef]
29. Hart, S.; Margheri, A.; Paci, F.; Sassone, V. Riskio: A serious game for cyber security awareness and education. *Comput. Secur.* **2020**, *95*, 101827. [CrossRef]
30. Jaffray, A.; Finn, C.; Nurse, J.R. Sherlocked: A detective-themed serious game for cyber security education. In Proceedings of the Human Aspects of Information Security and Assurance: 15th IFIP WG 11.12 International Symposium, HAISA 2021, Virtual Event, 7–9 July 2021; Proceedings 15. Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 35–45.
31. Gaurav, D.; Kaushik, Y.; Supraja, S.; Yadav, M.; Gupta, M.P.; Chaturvedi, M. Empirical study of adaptive serious games in enhancing learning outcome. *Int. J. Serious Games* **2022**, *9*, 27–42. [CrossRef]
32. Skinner, B.F. An operant analysis of problem solving. In *Problem Solving: Research, Method, and Theory*; Kleinmuntz, B., Ed.; Wiley: New York, NY, USA, 1966; pp. 225–257.
33. Skinner, B.F. An operant analysis of problem solving, Note 6.1–6.4. In *Contingencies of Reinforcement: A Theoretical Analysis*; Skinner, B.F., Ed.; Appleton-Century-Crofts: New York, NY, USA, 1969; pp. 157–171.
34. Pierce, W.D.; Cheney, C.D. *Behavior Analysis and Learning: A Biobehavioral Approach*; Routledge: London, UK, 2017.
35. Deterding, S.; Sicart, M.; Nacke, L.; O'Hara, K.; Dixon, D. Gamification. Using game-design elements in non-gaming contexts. In Proceedings of the CHI'11 Extended Abstracts on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 2425–2428.
36. Skinner, B.F. The shame of American education. *Am. Psychol.* **1984**, *39*, 947. [CrossRef]
37. Di Nocera, F.; Tempestini, G. Getting rid of the usability/security trade-off: A behavioral approach. *J. Cybersecur. Priv.* **2022**, *2*, 245–256. [CrossRef]

38. Utz, C.; Degeling, M.; Fahl, S.; Schaub, F.; Holz, T. (Un) informed consent: Studying GDPR consent notices in the field. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 973–990.
39. Giese, J.; Stabauer, M. Factors that Influence Cookie Acceptance. In *HCI in Business*; Fui-Hoon Nah, F., Siau, K., Eds.; Government and Organizations; HCII 2022; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13327.
40. Kulyk, O.; Hilt, A.; Gerber, N.; Volkamer, M. "This Website Uses Cookies": Users' Perceptions and Reactions to the Cookie Disclaimer. In Proceedings of the European Workshop on Usable Security (EuroUSEC), London, UK, 23 April 2018.
41. Bravo-Lillo, C.; Cranor, L.; Komanduri, S.; Schechter, S.; Sleeper, M. Harder to ignore? Revisiting {Pop-Up} fatigue and approaches to prevent it. In Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS 2014), Menlo Park, CA, USA, 9–11 July 2014; pp. 105–111.
42. Bravo-Lillo, C.; Komanduri, S.; Cranor, L.F.; Reeder, R.W.; Sleeper, M.; Downs, J.; Schechter, S. Your attention please: Designing security-decision UIs to make genuine risks harder to ignore. In Proceedings of the Ninth Symposium on Usable Privacy and Security, Newcastle, UK, 24–26 July 2013; pp. 1–12.
43. NordVPN Misfortune Cookie? Billions of Stolen Cookies Expose Your Data. Available online: https://nordvpn.com/research-lab/stolen-cookies-study/ (accessed on 1 July 2024).
44. Higley, E. Defining Young Adulthood. *DNP Qualif. Manuscr.* **2019**, *17*, 1–28. Available online: https://repository.usfca.edu/dnp_qualifying/17 (accessed on 1 July 2024).
45. Zhao, J.; Wang, G.; Dally, C.; Slovak, P.; Edbrooke-Childs, J.; Van Kleek, M.; Shadbolt, N. I make up a silly name' Understanding Children's Perception of Privacy Risks Online. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019; pp. 1–13.
46. Alanazi, M.; Freeman, M.; Tootell, H. Exploring the factors that influence the cybersecurity behaviors of young adults. *Comput. Hum. Behav.* **2022**, *136*, 107376. [CrossRef]
47. Razali, N.E.M.; Ramli, R.Z.; Mohamed, H.; Zin NA, M.; Rosdi, F.; Diah, N.M. Identifying and validating game design elements in serious game guidelines for climate change. *Heliyon* **2022**, *8*, e08773. [CrossRef]
48. Starks, K. Cognitive behavioral game design: A unified model for designing serious games. *Front. Psychol.* **2014**, *5*, 28. [CrossRef]
49. Antin, J.; Churchill, E.F. Badges in Social Media: A Social Psychological Perspective. In Proceedings of the CHI 2011 Gamification Workshop Proceedings, Vancouver, BC, Canada, 7–12 May 2011; ACM Press: New York, NY, USA, 2011.
50. Isbister, K.; Schaffer, N. *Game Usability: Advancing the Player Experience*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2008.
51. Lankoski, P.; Björk, S. Gameplay design patterns for believable non-player characters. In Proceedings of the DiGRA 2007 Conference: Situated Play, Tokyo, Japan, 24–28 September 2007.
52. Butler, A.C.; Roediger, H.L., III. Testing improves long-term retention in a simulated classroom setting. *Eur. J. Cogn. Psychol.* **2007**, *19*, 514–527. [CrossRef]
53. Blunt, R. Do serious games work? Results from three studies. *ELearn* **2009**, *2009*, 1. [CrossRef]
54. Adachi, P.J.; Willoughby, T. More than just fun and games: The longitudinal relationships between strategic video games, self-reported problem solving skills, and academic grades. *J. Youth Adolesc.* **2013**, *42*, 1041–1052. [CrossRef] [PubMed]
55. Rogers, S. *Level Up! The Guide to Great Video Game Design*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
56. Kuntjara, H.; Almanfaluthi, B. Character design in games analysis of character design theory. *J. Games Game Art Gamification* **2017**, *2*, 42–47. [CrossRef]
57. Burgerman, J. 20 Top Character Design Tips. Available online: http://www.creativebloq.com/character-design/tips-51326432015 (accessed on 7 August 2024).
58. Ivy, J.W.; Meindl, J.N.; Overley, E.; Robson, K.M. Token economy: A systematic review of procedural descriptions. *Behav. Modif.* **2017**, *41*, 708–737. [CrossRef] [PubMed]
59. Vorderer, P.; Klimmt, C.; Ritterfeld, U. Enjoyment: At the heart of media entertainment. *Commun. Theory* **2004**, *14*, 388–408. [CrossRef]
60. Busse, J.; Lange, A.; Hobert, S.; Schumann, M. How to Design Learning Applications That Support Learners in Their Moment of Need–Didactic Requirements of Micro Learning. In Proceedings of the Americas Conference on Information Systems (AMCIS 2020), Salt Lake City, UT, USA, 12–16 August 2020; pp. 1–10.
61. Leong, K.; Sung, A.; Au, R.; Lee, C. A study of preferred learning time of online learners in multimedia microlearning in higher education contexts. *Online J. TVET Pract.* **2022**, *7*, 11–23. [CrossRef]

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Information Editorial Office
E-mail: information@mdpi.com
www.mdpi.com/journal/information

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-7469-9