

Article

Synthetic Data Augmentation for Robust Classification of Diabetic vs. Non-Diabetic Blood FTIR Spectra

Ahmed Fadlelmoula ^{1,2}, Kirill N. Boldyrev ³, Margarida Gonçalves ¹, Helena Torres ², Susana O. Catarino ^{1,4}, Graça Minas ^{1,4} and Vitor Carvalho ^{2,*}

¹ Center for MicroElectromechanical Systems (CMEMS-UMinho), University of Minho, 4800-058 Guimarães, Portugal; id9247@alunos.uminho.pt (A.F.); id11729@alunos.uminho.pt (M.G.); scatarino@dei.uminho.pt (S.O.C.); gminas@dei.uminho.pt (G.M.)

² 2Ai, School of Technology, Polytechnic University of Cávado and Ave, 4750-810 Barcelos, Portugal; htorres@ipca.pt

³ Beijing Institute of Technology (BIT), Zhuhai BIT, Zhuhai 519088, China; kn.boldyrev@bit.edu.cn

⁴ LABBELS—Associate Laboratory, 4710-057 Braga, Portugal

* Correspondence: vcarvalho@ipca.pt (V.C.)

Abstract

Early detection of diabetes mellitus (DM) is essential for preventing disease progression and improving clinical outcomes. However, developing robust machine learning (ML) models for diabetes diagnosis is often constrained by limited data availability, privacy regulations, and challenges with data sharing. This study investigates a privacy-preserving synthetic data augmentation framework for classifying diabetic and non-diabetic blood serum samples using Fourier Transform Infrared (FTIR) spectroscopy. Two deep generative approaches, Autoencoders (AEs) and Generative Adversarial Networks (GANs), were evaluated for their ability to generate realistic synthetic FTIR spectra while preserving the statistical and biochemical characteristics of the original dataset. Synthetic datasets generated by the AE and GAN models were assessed using six ML classifiers: Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting (GB), Logistic Regression (LoR), and Decision Tree (DT). Model performance was evaluated using accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC) curves, and Area Under the Curve (AUC). Results showed that AE-generated spectra retained stronger discriminative characteristics and were more easily distinguished from the original spectra, whereas GAN-generated spectra exhibited lower classifier separability, suggesting closer alignment with the original data distribution and greater realism for privacy-oriented data augmentation. Correlation analysis demonstrated high spectral fidelity for both approaches. Compared with the original spectra, AE-generated spectra achieved $r = 0.9990$ and $R^2 = 0.9999$, whereas GAN-generated spectra achieved $r = 0.9982$ and $R^2 = 0.9965$. The most prominent diabetes related spectral variations were observed in the carbohydrate ($1000\text{--}1200\text{ cm}^{-1}$), Amide I ($\sim 1650\text{ cm}^{-1}$), and lipid-associated ($3000\text{--}3500\text{ cm}^{-1}$) regions. To explore the transferability of the proposed framework, a preliminary experimental feasibility study was conducted using independently acquired whole blood FTIR spectra. The generated spectra showed strong agreement with the measured whole blood spectra, demonstrating the potential applicability of the framework under alternative sampling conditions. Because the experimental cohort included only one diabetic volunteer, this analysis was intended solely as a proof-of-concept assessment of spectral feasibility and methodological transferability, rather than as a validation of diabetes classification performance. Overall, the findings demonstrate that synthetic data generation can effectively augment limited

Academic Editors: Zhou Yang and Fang Jin

Received: 22 May 2026

Revised: 26 June 2026

Accepted: 26 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

FTIR datasets while preserving privacy and key spectral characteristics. The proposed framework provides a promising foundation for privacy-aware biomedical data augmentation and future development of robust FTIR diabetes screening systems. The results should be interpreted as methodological evidence of feasibility and synthetic data utility rather than as evidence of clinical diagnostic readiness, as the serum dataset remains modest in size and the independent whole-blood experiment was intentionally exploring.

Keywords: Autoencoders; diabetes classification; FTIR spectroscopy; generative adversarial networks; synthetic data generation

1. Introduction

Disease diagnosis involves identifying the specific pathological condition responsible for a patient's symptoms [1]. However, because many clinical manifestations are nonspecific and may overlap across multiple conditions, achieving an accurate diagnosis often presents a considerable challenge [2]. Accurate disease identification is a fundamental prerequisite for effective treatment and patient management [3]. Recent advances in machine learning (ML), a key subfield of artificial intelligence (AI), have provided promising approaches for addressing these diagnostic challenges [4]. ML techniques facilitate the early detection and diagnosis of a wide range of diseases, including cancer, diabetes, cardiovascular disorders, and dermatological conditions [5]. By enhancing diagnostic speed, accuracy, and reliability, these approaches significantly strengthen clinical decision-making processes [6]. Numerous ML models and frameworks have been proposed to improve diagnostic efficiency and precision [7]. Unlike traditional rule-based systems, ML algorithms autonomously learn from data and construct predictive models that support early-stage diagnosis and provide data-driven clinical insights [8]. Consequently, ML diagnostic tools not only enhance diagnostic accuracy and efficiency but also reduce diagnostic time and improve predictive outcomes for various infectious and chronic diseases [9]. Training ML algorithms requires access to real-world patient data to enable models to learn underlying patterns and effectively distinguish between different clinical conditions. However, the acquisition and utilization of large-scale healthcare data present numerous challenges that hinder accessibility and scalability. One of the most significant obstacles is data privacy, which refers to the protection of sensitive information, particularly personally identifiable healthcare data [10]. Data privacy emphasizes appropriate data governance practices, including policies, authorization mechanisms, and regulatory frameworks that ensure patient information is collected, shared, and utilized responsibly [11]. Another critical concern is data security, which focuses on safeguarding information from unauthorized access, malicious attacks, and data breaches [12]. While data security plays an essential role in protecting information integrity and availability, it alone does not fully address privacy-related challenges. In the healthcare sector, protecting sensitive patient information, including medical records, personal identifiers, and other health-related data, is paramount [13]. Personal data refers to information capable of identifying an individual, whereas medical data encompasses information generated during healthcare delivery, such as diagnostic results, treatment histories, and clinical outcomes [13,14]. Furthermore, healthcare data are often fragmented across different electronic health record (EHR) systems and institutions, complicating efforts to access, integrate, and standardize information for clinical and research purposes [15]. Differences in data definitions, coding schemes, and storage formats further obstruct compatibility and data harmonization. Additional

challenges, including imbalanced datasets, labeling uncertainties, data overload, and inconsistencies in data collection, represent significant limitations in healthcare data management [15].

Beyond privacy preservation and data accessibility, the development of reliable AI systems for healthcare faces significant challenges in external validation, reproducibility, robustness, and clinical translation [6]. ML models trained on data from a single institution or a limited patient population often perform worse when applied to independent cohorts due to variations in demographics, clinical practices, instrumentation, and data acquisition protocols [16]. Consequently, external validation using independent datasets has become essential for establishing the robustness, generalizability, and clinical applicability of biomedical AI models [17]. However, large-scale external validation studies remain difficult because access to clinical datasets is frequently constrained by privacy regulations, institutional policies, and ethical considerations [18]. These restrictions also hinder scientific reproducibility, as independent researchers may be unable to access the original datasets required to verify published findings, benchmark alternative methodologies, or assess model robustness across diverse populations [19]. As a result, many AI models that achieve promising performance in controlled research environments face challenges when translated into real-world clinical practice.

Synthetic data generation has emerged as a promising strategy for addressing these limitations. By generating artificial datasets that preserve the statistical, structural, and biochemical characteristics of real-world data while minimizing privacy risks, synthetic data can facilitate secure data sharing, support multi-institutional validation studies, and improve the reproducibility of biomedical research. Furthermore, synthetic data augmentation can increase dataset size and diversity, reduce the impact of class imbalance, and improve model robustness when real-world data are scarce or underrepresented [20]. These advantages are particularly relevant in healthcare, where access to large, high-quality datasets is often restricted. In the context of FTIR disease screening, synthetic data generation may therefore contribute not only to dataset augmentation but also to the development of clinically reliable, reproducible, and generalizable diagnostic systems capable of supporting real-world implementation [21].

Successful deployment of AI systems in healthcare also requires consideration of operational factors, including integration into existing clinical workflows, regulatory compliance, model interpretability, and long-term performance monitoring. Models that perform well under controlled research conditions may degrade when exposed to new patient populations, changing clinical practices, or variations in instrumentation and data-acquisition procedures [22]. Consequently, continuous validation, periodic model updating, and prospective multi-center evaluation are essential to ensure reliable deployment in real-world healthcare and public health settings. Although the present study focuses on methodological validation using FTIR blood serum spectra, future work should investigate these operational aspects to support the translation of synthetic data-enabled AI systems into routine clinical practice [23].

These challenges are particularly relevant to FTIR diabetes classification, where available datasets are often limited in size, collected from single institutions, and acquired using institution-specific measurement protocols [24]. Consequently, ML models developed using a single dataset may exhibit reduced robustness and generalizability when applied to independent populations, different instrumentation, or alternative acquisition settings [25]. In this context, synthetic data augmentation may provide a practical mechanism for expanding training datasets, increasing data diversity, supporting reproducibility studies, and facilitating preliminary external validation while preserving patient privacy [26]. Although synthetic data cannot replace large-scale prospective clinical validation, it can contribute to the development of more robust and generalizable AI models by reducing de-

pendence on restricted clinical datasets and enabling broader evaluation of model performance across diverse scenarios [27]. In addition, synthetic datasets provide a valuable intermediary resource for benchmarking algorithms, conducting preliminary external validation studies, and evaluating model robustness across simulated population variability before large-scale clinical deployment [28]. Cost efficiency may also be achieved by reducing reliance on expensive and difficult-to-obtain medical data [29].

ML synthetic data generation involves training models on original datasets to learn their underlying patterns, distributions, and dependencies. Once adequately trained, these models can generate realistic synthetic samples that closely resemble the original data while preserving privacy [30]. For example, in diabetes research, ML models can generate synthetic patient profiles containing attributes such as age, blood glucose levels, insulin concentrations, and health outcomes that appear realistic but do not correspond to any actual individual. Commonly adopted techniques for synthetic data generation include Generative Adversarial Networks (GANs), in which a generator and discriminator compete to produce highly realistic synthetic data, and Autoencoders (AEs), which learn compressed data representations and reconstruct data distributions before generating new samples. Additionally, agent and simulation approaches enhanced with ML can be used to model realistic physiological behaviors and medical events [31]. The application of synthetic data generation has emerged as a promising approach for the early detection and analysis of various diseases, including Diabetes Mellitus (DM), the focus of the present study [32]. Synthetic data generation refers to the process of creating artificial datasets that replicate the statistical and structural properties of real-world data without being derived from actual patients, events, or systems [33].

Although alternative approaches such as transfer learning and cross-institutional data sharing have been proposed to mitigate data scarcity, their application in FTIR healthcare research remains challenging [34]. Transfer learning requires sufficiently large and domain-relevant source datasets, which are currently limited in this field, whereas cross-institutional data sharing is frequently restricted by privacy regulations, institutional policies, and differences in data acquisition protocols [35]. Furthermore, although transfer learning has achieved considerable success in image analysis applications, comparable large-scale FTIR blood serum repositories suitable for transfer learning remain scarce. Consequently, transferring representations learned from unrelated spectral or biomedical datasets may not adequately capture disease-specific biochemical signatures relevant to FTIR diabetes classification. In contrast, synthetic data generation directly learns from the target FTIR dataset and can augment disease-specific spectral variability while preserving privacy and meeting data governance requirements. Therefore, synthetic data augmentation was selected as the primary strategy in this study because it simultaneously addresses dataset scarcity, privacy preservation, class balancing, reproducibility, and model development requirements [36,37].

The limited availability of large-scale FTIR blood serum datasets represents an additional challenge for the development of robust ML models [38]. In biomedical spectroscopy, dataset collection is often constrained by patient recruitment requirements, ethical approval procedures, informed consent regulations, sample acquisition costs, and limited access to specialized instrumentation [39]. Consequently, many published FTIR studies on diabetes have relied on relatively small cohorts. For example, Li et al. analyzed 47 blood samples [40], whereas Guang et al. evaluated 113 subjects (51 diabetic and 62 healthy) [41]. These examples illustrate that datasets in this field often consist of only a few dozen to approximately 100 participants. The dataset used in the present study (100 FTIR spectra) is therefore comparable in size to those commonly reported in the FTIR diabetes literature, further motivating the investigation of synthetic data augmentation as a mechanism to increase data diversity and support more robust ML model development. Nevertheless,

such limited sample sizes may restrict model generalizability and increase the risk of over-fitting.

Accordingly, this study investigates whether synthetic data generation can support privacy-compliant ML analysis of FTIR blood serum spectra for discrimination between diabetic and non-diabetic samples. Specifically, the study presents a comparative assessment of AE- and GAN-generated synthetic data across three complementary dimensions: preservation of spectral fidelity, utility for downstream classification, and realism-oriented augmentation. DM was selected as the case study because it is a clinically relevant chronic metabolic disorder for which early detection is essential to reduce long-term complications and improve patient outcomes [42,43]. In addition to the primary serum analysis, this study includes a preliminary experimental validation using independently acquired whole blood FTIR spectra. This validation was conducted as an initial feasibility assessment to determine whether the proposed framework can be applied to independently acquired whole-blood FTIR spectra under a more application-oriented sampling condition. The objective of this preliminary experiment was to assess spectral feasibility and methodological transferability rather than diabetes classification performance. Therefore, the novelty of this work lies not only in the comparative evaluation of AE- and GAN-based synthetic FTIR data augmentation for diabetes classification, but also in the preliminary demonstration of its potential translational applicability beyond the original serum dataset.

The paper is organized into seven sections: Section 2 reviews the state of the art in synthetic data generation and FTIR diabetes diagnosis, highlighting recent ML approaches and their applications in healthcare. Section 3 presents the theoretical foundations of synthetic data generation, including the architectures, learning mechanisms, and training dynamics of AE and GAN. Section 4 describes the proposed methodology, encompassing the synthetic data generation workflow, data preprocessing procedures, dimensionality reduction, dataset partitioning, and validation strategy. Section 5 details the dataset characteristics, computational environment, model architectures, hyperparameter optimization process, classifier evaluation framework, and experimental setup. Section 6 presents and discusses the experimental results, including loss-function analysis, synthetic-data realism assessment, classifier performance evaluation, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) analyses, correlation analysis, coverage assessment, and preliminary experimental validation using independently acquired whole-blood FTIR spectra. Finally, Section 7 summarizes the main findings, discusses the study's limitations, and outlines directions for future research and clinical translation.

2. Related Work

This section reviews the state of the art in the application of ML techniques for synthetic tabular data generation, with particular emphasis on methods for generating and distinguishing between original and synthetic data. The discussion is structured into two main parts: the first addresses existing approaches to synthetic data generation, while the second explores methodologies applied to diabetes diagnosis.

2.1. Synthetic Data Methods

Endres et al. (2022) [44] conducted a comprehensive comparative analysis of seven synthetic data generation models: Synthetic Minority Oversampling Technique (SMOTE), GAN, AE, Data Synthesizer, Synthetic Data Vault (SDV)-Gaussian Copula, SDV-GAN, and SynthPop Non-Parametric, using three open-source datasets: adult Census, Airbnb, and Airlines. The study evaluated each method based on processing time, data quality, and the preservation of pairwise correlations between original and synthetic datasets. Evaluation metrics included pairwise correlation, proximity, and SDMetrics, a dataset ag-

nostic framework incorporating statistical, detection, and efficacy measures. The findings indicated that while deep generative models such as GAN and AE produced high-fidelity outputs, simpler techniques like the Gaussian Copula and SynthPop achieved competitive performance with significantly greater computational efficiency. Furthermore, the study discussed key implementation challenges related to data preprocessing, model training, and scalability across heterogeneous datasets. Overall, the paper provided an empirical foundation for selecting appropriate synthetic data generation techniques based on dataset characteristics, desired fidelity, and computational constraints [44].

Similarly, Miura et al. (2024) [45] empirically evaluated various synthetic data generation techniques applied to real-world medical datasets, specifically utilizing the Diagnosis Procedure Combination (DPC) dataset from Ehime University Hospital. Their study compared three methodological classes—statistical-based, graphical model, and deep neural network approaches—under differential privacy constraints. Evaluation criteria included statistical distribution distances, ML performance, and correlation matrix similarity between original and synthetic data. The experimental results demonstrated that statistical and graphical model approaches, particularly the Gaussian Copula and Adaptive Inference Model (AIM), outperformed deep neural network models such as the Conditional Tabular Generative Adversarial Network (CTGAN) in maintaining data utility when differential privacy was enforced. The authors concluded that while differential privacy enhances privacy protection, it inevitably introduces varying degrees of data quality degradation depending on the synthesis method. This study offered practical insights into the trade-offs between privacy and utility in synthetic healthcare data generation, emphasizing the effectiveness of combining statistical modelling with differential privacy for medical research applications [45].

In another contribution, Rashidi et al. (2024) [46] introduced the Synthetic Tabular Neural Generator (STNG), a fully automated framework designed for generating and validating synthetic tabular data to address healthcare data accessibility challenges under the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). Their findings showed that STNG's multi-functional variants consistently outperformed conventional synthetic data generation methods in seven out of nine binary datasets and all multiclass datasets evaluated. STNG thus represents a scalable, regulation-compliant framework that supports automated synthetic data generation and validation in healthcare, promoting reproducibility, privacy preservation, and the accessibility of high-quality tabular datasets [46].

More recently, Bae et al. (2025) [47] addressed the issue of class imbalance in small, data-scarce medical datasets by proposing an incremental synthetic data generation system designed to improve data quality over multiple iterations. This iterative approach enhances model sensitivity and generalization while mitigating overfitting, thereby improving performance in imbalanced, small-scale medical datasets. The system supports individualized health risk prediction, such as asthma attack forecasting, by iteratively refining synthetic data generation based on model feedback. This work establishes a novel, data-centric paradigm that extends traditional oversampling methods through performance-driven, iterative refinement to produce higher quality synthetic data for rare event prediction and personalized medicine [47].

2.2. Diabetes Diagnosis Methods

Yang et al. (2018) [48] developed a rapid and non-invasive diagnostic approach for prediabetes based on Fourier Transform Mid Infrared Attenuated Total Reflection (FTIR-ATR) spectroscopy of peripheral blood samples. The study integrated Classification and Regression Trees (CART) and Extreme Gradient Boosting (XGBoost) ML algorithms for disease classification. Following Savitzky–Golay smoothing and Principal Component

Analysis (PCA) for data preprocessing and feature extraction, the XGBoost model demonstrated superior diagnostic performance, achieving 93.33% accuracy, 100% specificity, and 85% sensitivity, outperforming the CART model. This method eliminates the need for chemical reagents, offering a fast, reliable, and scalable solution for large-scale prediabetes screening.

Li et al. (2025) [40] investigated a non-invasive approach for blood glucose measurement based on infrared (IR) spectroscopy coupled with Partial Least Squares (PLS) regression. In their study, 47 blood samples were collected, and glucose concentrations were determined using a standard glucometer for reference. The spectral data were subsequently analysed using an FTIR spectrometer, and the developed model achieved a mean estimation error of 7 mg/dL, comparable to the performance of commercial invasive glucose monitoring devices. The findings demonstrate that the integration of IR spectroscopy with Partial Least Squares (PLS) regression provides a promising framework for non-invasive glucose estimation and could potentially be extended to the quantification of other biochemical constituents in blood.

Guang et al. (2020) [41] investigated the application of FTIR-ATR spectroscopy of whole blood in combination with the Extreme Gradient/Boosting (XGBoost) algorithm for the diagnosis of type 2 diabetes. In this study, blood samples from 51 diabetic patients and 62 healthy controls were analysed. Preprocessing techniques, including Savitzky–Golay smoothing and PCA, were employed to enhance spectral quality and improve classification performance. The optimized XGBoost model achieved 95.24% sensitivity, 96.00% specificity, and 95.65% overall accuracy, highlighting the potential of FTIR-ATR spectroscopy as a rapid, reagent-free, and reliable diagnostic tool for diabetes detection.

Dharmarathne et al. (2024) [49] employed ATR-FTIR spectroscopy combined with chemometric modelling for the rapid screening of diabetes using human blood serum. Serum samples obtained from diabetic and non-diabetic subjects were analysed, and multivariate statistical techniques, including PCA and Partial Least Squares–Discriminant Analysis (PLS-DA), were utilized to classify samples based on spectral biomarkers. The integrated ATR-FTIR and chemometric approach demonstrated high accuracy, sensitivity, and specificity, underscoring its potential as a fast, non-invasive, and reagent-free diagnostic tool for diabetes detection and screening.

These studies are closely aligned with recent advancements in non-invasive diabetes diagnosis using FTIR spectroscopy integrated with ML techniques. The proposed approach addresses the challenges associated with limited data availability and privacy-sensitive clinical datasets by incorporating synthetic data generation through AE and GAN to augment FTIR blood serum datasets. This framework establishes a foundation for the development of more robust, generalizable, and ethically compliant ML models for reliable diabetes detection.

3. Synthetic Data Generation Theoretical Foundations

The following subsections outline the theoretical foundations and architectural principles of AEs and GANs, which constitute the methodological basis for the synthetic data generation framework employed in this study.

3.1. Autoencoder Architecture and Learning Process

An AE is an artificial neural network designed to learn data representations in an unsupervised manner [50]. Its primary objective is to capture the most significant features of high-dimensional data by compressing it into a lower-dimensional representation, commonly referred to as an encoding [50]. This process enables dimensionality reduction and efficient data representation without requiring labeled inputs. AE accomplishes this through a two-part architecture consisting of an encoder and a decoder [50]. The encoder

(green) compresses the input data into a compact latent space, while the decoder (blue) reconstructs the original input from this compressed representation. By minimizing reconstruction errors during training, the network learns to identify and preserve the essential features of the data [51]. The general architecture of an AE comprises three key components, as illustrated in Figure 1. The encoder begins with an input layer that receives raw data. Through successive hidden layers, the data dimensionality is reduced, allowing for the capture of important patterns and characteristics.

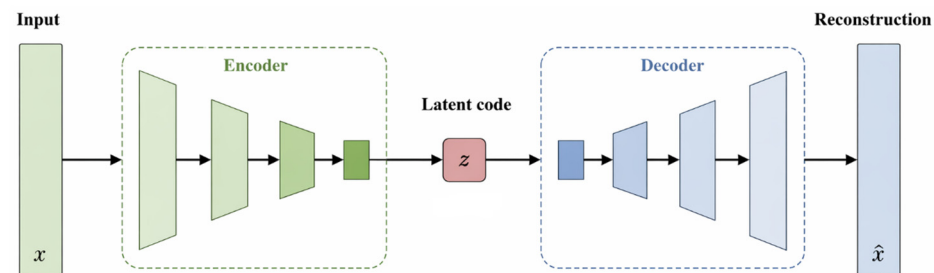


Figure 1. AE architecture.

This process culminates in the bottleneck layer, which represents the compressed latent encoding of the input data [51]. The latent code serves as the core of the network, encapsulating compact knowledge representations that capture the most salient features of the data [52]. The decoder reconstructs the original input from this latent representation by progressively increasing the dimensionality of the hidden layers until reaching the output layer. Ideally, the reconstructed output closely approximates the original input [51]. Training AE involves minimizing a reconstruction loss, which quantifies the discrepancy between the original input and its reconstruction [52]. Commonly used loss functions include Mean Squared Error (MSE) for continuous data and binary cross-entropy for binary data [53]. Once trained, the encoder can operate independently to generate compact representations of new data that resemble those in the training set. To prevent overfitting and encourage meaningful feature extraction, several constraints are typically imposed. Limiting the number and size of hidden layers compels the model to learn only the most representative features [54], while incorporating regularization terms in the loss function discourages trivial solutions and enhances generalization [55]. Additionally, applying denoising techniques by introducing noise into the input data and training the model to reconstruct the clean version of the data improves robustness [56]. Adjusting activation functions to maintain sparsity, ensuring that most neurons remain inactive, further reduces model complexity and promotes efficient representation learning [57]. Through these mechanisms, AE learn compact, meaningful, and generalizable data representations that are widely used in dimensionality reduction, feature extraction, and anomaly detection.

3.2. Generative Adversarial Network Architecture and Training Dynamics

A GAN consists of two distinct models, the generator and the discriminator, trained using adversarial learning methods [58]. The first model is a generative model that attempts to generate new input samples that are completely different from the real inputs. The generator has never been exposed to real input; however, it operates with an input structure similar to that of the original dataset. The generator accepts latent vectors (random noise) as input and produces new samples that align with the distribution of original information samples [58]. The second model is the discriminator, whose primary function is to differentiate between data and estimate the likelihood that it originates from either the original dataset or the generator [58]. Figure 2 illustrates the GAN model, which be-

gins with a latent vector (represented by blue dots on the left). This vector is a random noise sample, typically drawn from a simple distribution such as a Gaussian or uniform distribution. It represents the seed from which the generator creates new data. The generator (blue/purple box) receives the latent vector and transforms it through a neural network into synthetic (fake) data samples. These fake samples are meant to imitate the distribution of original samples (pink box). Both original samples (from the training dataset) and fake samples (produced by the generator) are fed into the discriminator (green box). The discriminator's job is to classify each input as real or fake. The discriminator's output is the probability of the data being original or not. Based on its prediction, a discriminator loss is computed to measure how well it correctly identifies real vs. fake data. If it misclassifies, the loss increases, prompting it to adjust its weight during backpropagation. The generator's goal is to produce fake data that the discriminator classifies as real. Hence, the generator loss depends on the discriminator's output when it is fooled (i.e., when fake data are classified as real). The generator updates its parameters based on this feedback to create more realistic data in the next iteration. The GAN training process alternates between training the discriminator, keeping the generator fixed to better distinguish between real and fake data, and training the generator, keeping the discriminator fixed to produce more realistic outputs. Over time, the generator learns to produce data that closely mimics the original distribution, and the discriminator becomes less confident because fake and real data become increasingly indistinguishable, until the generator produces highly realistic data. At this point, the discriminator can no longer distinguish real from fake with greater than random guessing accuracy (50%) [58]. GANs can learn complex, high-dimensional data distributions and generate highly realistic synthetic data, making them particularly powerful in fields such as healthcare, where data availability and privacy pose significant challenges.

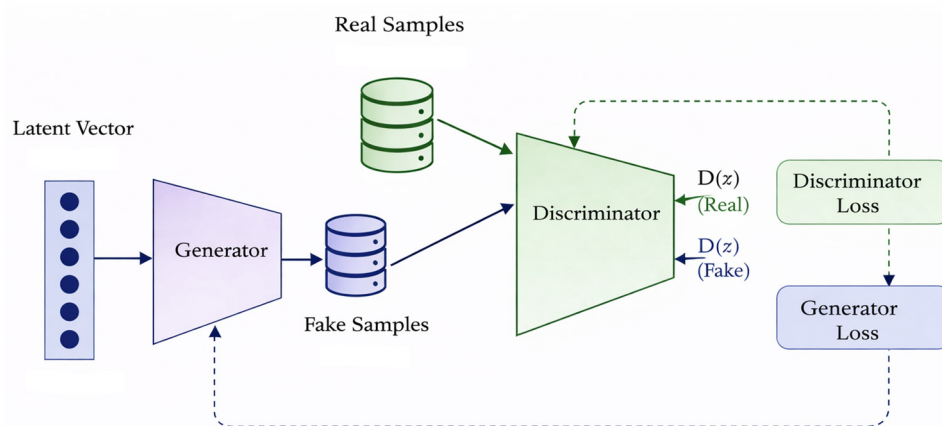


Figure 2. GAN Architecture.

4. Methodology

This section presents the framework used to generate synthetic data from the collected blood sample dataset. It further describes the methodologies used in the data generation process, outlines the evaluation strategy for assessing the quality of the synthesized data, and details the procedures for experimental data preparation and analysis.

4.1. Synthetic Data Generation Workflow

The proposed methodology is structured into two principal phases: (i) synthetic data generation and (ii) evaluation of the generated data. This framework was designed to ensure that the synthesized FTIR spectra accurately capture the statistical and biochemical characteristics of the original dataset while remaining suitable for downstream predictive

tasks such as diabetes classification. As illustrated in Figure 3a, the overall pipeline integrates data preprocessing, generative modeling, iterative training, and classifier evaluation. Figure 3b further decomposes this workflow into a sequence of well-defined stages, thereby ensuring methodological clarity and reproducibility.

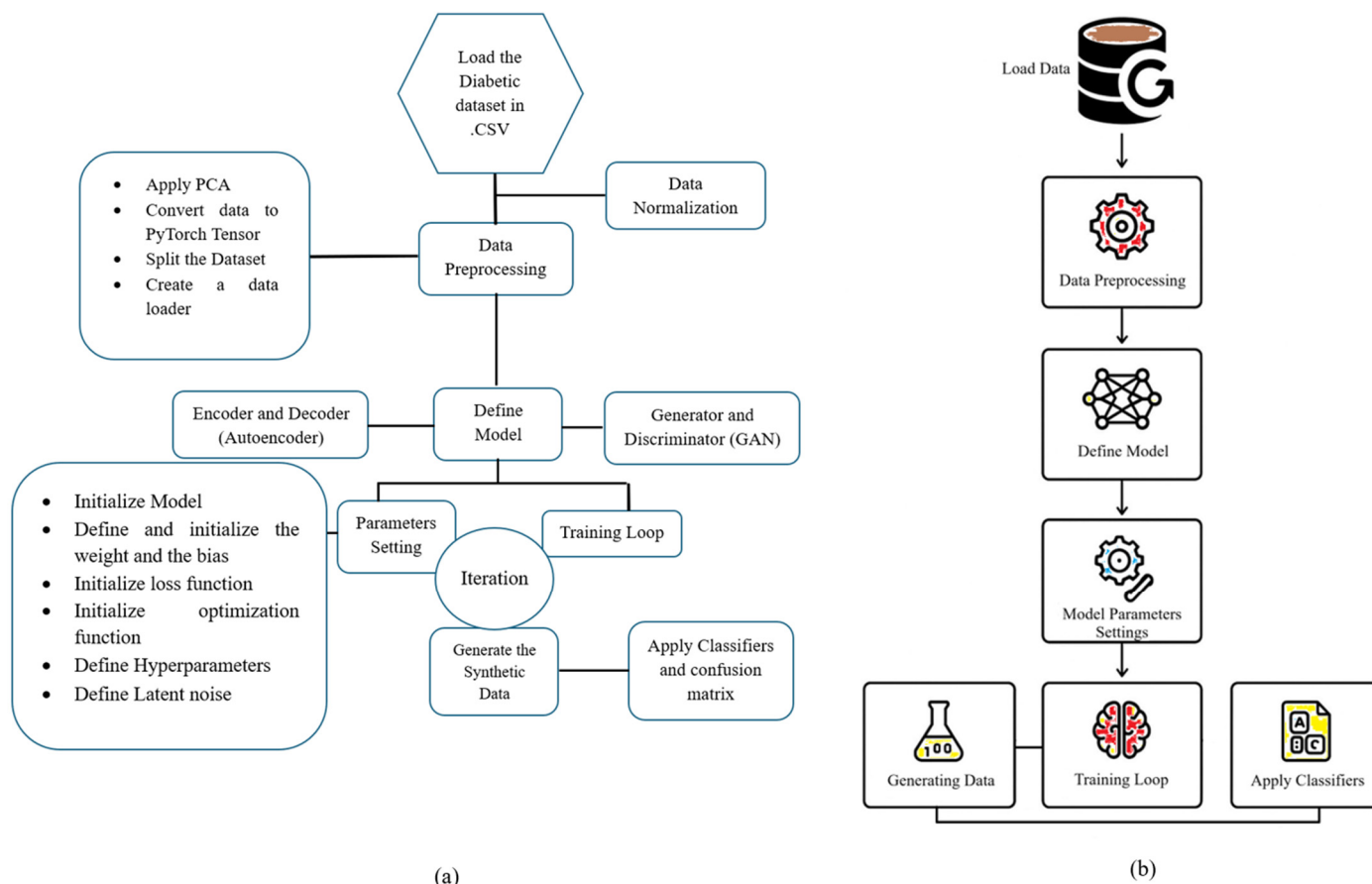


Figure 3. Steps required to generate synthetic data: (a) methodology pipeline; (b) flowchart of the synthetic data generation pipeline.

4.1.1. Data Loading and Preprocessing

The process begins by loading the original FTIR blood serum dataset, containing both diabetic and non-diabetic spectra, from a structured CSV file. During synthetic data generation, the AE and GAN models were trained separately within each disease class to preserve class-specific spectral characteristics and maintain the original class distribution. Preprocessing was subsequently applied to improve data quality and model compatibility. Feature scaling was performed using normalization techniques, including StandardScaler and MinMaxScaler [59], to mitigate disparities in feature magnitudes and ensure stable model convergence. To address the high dimensionality of the spectral data, PCA was applied as a dimensionality reduction technique. By retaining 99.9% of the variance, PCA reduced noise and redundancy while preserving the most informative spectral features. The use of PCA prior to generative modeling was intended to reduce computational complexity, mitigate overfitting associated with the limited dataset size, and improve the stability of AE and GAN training by removing highly correlated spectral variables. The transformed data were subsequently used as inputs to the generative models. The preprocessing workflow therefore consisted of five sequential stages: (i) spectral normalization using StandardScaler (AE) or MinMaxScaler (GAN), (ii) dimensionality reduction using PCA with 99.9% variance retention, (iii) conversion of the processed spectra into PyTorch

tensor representations, (iv) partitioning of the dataset into training, validation, and testing subsets, and (v) training of the generative models using the processed feature space. Following preprocessing, the dataset was partitioned into training, validation, and test subsets, as described further in Section 4.3. The processed data were then converted into tensor representations compatible with PyTorch, and DataLoaders were constructed to facilitate efficient mini-batch processing during model training. Details of the software versions used in this study are provided in Section 5.

4.1.2. Model Definition and Parameter Configuration

Two generative modeling approaches were employed: AE and GAN. The AE architecture consists of an encoder and a decoder, which learn compressed latent representations of the input data and subsequently reconstruct the original spectra. In contrast, the GAN framework comprises a generator and a discriminator trained in an adversarial manner, where the generator attempts to produce realistic synthetic spectra and the discriminator attempts to distinguish between original and generated samples. Prior to training, model parameters, including network weights and biases, were initialized. For the GAN model, Binary Cross-Entropy Loss (BCELoss) was used to quantify the adversarial objective, whereas Mean Squared Error Loss (MSELoss) was employed for the AE model to measure reconstruction error. Optimization was performed using the Adam (Adaptive Moment Estimation) optimizer, which provides adaptive learning rates and improved convergence stability. Key hyperparameters, including learning rate, batch size, number of epochs, and latent-space dimensionality, were defined experimentally.

4.1.3. Training Procedure

The training process was conducted iteratively over multiple epochs using mini-batch gradient descent. For the GAN model, training followed an alternating optimization strategy. The discriminator was first updated to correctly classify original and generated samples, after which the generator was updated to improve its ability to produce realistic spectra. To enhance training stability, techniques such as label smoothing, noise injection, and multiple generator updates per iteration were incorporated. For the AE model, training involved forward propagation through the encoder–decoder architecture, computing the reconstruction loss, and parameter updates via backpropagation. Training and validation losses were monitored throughout the process to assess convergence and detect potential overfitting. This iterative learning procedure enabled the models to capture complex spectral patterns present in the original FTIR dataset and generate synthetic spectra with similar statistical and structural characteristics. Multiple training iterations across successive epochs ensured convergence and stable model performance.

4.1.4. Synthetic Data Generation and Evaluation

Upon completion of training, the generator (GAN) or decoder (AE) was used to produce synthetic datasets. These generated spectra were intended to replicate the statistical distribution and feature relationships of the original FTIR data. Following synthetic data generation, two complementary evaluation tasks were performed. The first focused on synthetic-data realism, where classifiers were trained to distinguish between original and synthetic spectra. Within this framework, low classifier separability indicates that the synthetic samples closely resemble the original data distribution and are therefore more realistic. The second evaluation focused on diabetes discrimination, in which classifiers were used to distinguish diabetic from non-diabetic samples in both the original and augmented datasets. These two evaluation tasks assess fundamentally different aspects of model performance and were therefore interpreted independently throughout the study.

To quantitatively assess the similarity between original and synthetic FTIR spectra, Pearson's correlation coefficient (r) and the coefficient of determination (R^2) were calculated—as presented further in Section 6.1.7. Pearson's correlation coefficient measures the strength of the linear relationship between the original and generated spectra, whereas R^2 quantifies the proportion of variance in the original spectra explained by the synthetic spectra. These metrics were computed according to Equations (1) and (2). The Pearson Correlation Coefficient is calculated using Equation (1):

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where x_i and y_i represent the absorbance values of the original and synthetic spectra, respectively, and $\bar{x}$ and $\bar{y}$ denote their corresponding mean values. The Coefficient of Determination R^2 was used to evaluate how well the synthetic spectra reproduce the variability of the original spectra. Values closer to 1 indicate a stronger agreement between the compared datasets. The coefficient is calculated using Equation (2):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

where y_i represents the original spectral values, $\hat{y}_i$ denotes the corresponding synthetic spectral values, and $\bar{y}$ is the mean of the original spectral values.

Finally, ML classifiers, including Support Vector Machines (SVM) and Random Forests (RF), were applied to assess the usability and predictive value of the generated data. Confusion-matrix analysis was subsequently used to compare predicted and actual labels, thereby evaluating whether the synthetic spectra preserved predictive characteristics comparable to those of the original FTIR dataset.

4.2. Data Preprocessing

PCA is a widely used statistical technique for dimensionality reduction and feature extraction, particularly in high-dimensional datasets such as those in spectroscopy. Before training the generative models, the FTIR spectra underwent model-specific normalization procedures to improve numerical stability and training performance. For the AE model, spectral variables were standardized using StandardScaler, which centers each feature and scales it to unit variance. This scaling approach is commonly employed in neural network reconstruction because it centers the data distribution and facilitates stable gradient propagation during training. For the GAN model, MinMaxScaler was applied to scale all spectral variables to the range [0, 1]. This normalization strategy was selected because the GAN generator used output activation functions that perform best within a bounded numerical range, thereby improving adversarial training stability and convergence. Different normalization methods were adopted because preliminary experiments demonstrated improved convergence and training stability when StandardScaler was used for the AE and MinMaxScaler was used for the GAN.

Following normalization, PCA was applied to the spectra to reduce redundancy and noise while preserving the most informative biochemical features. PCA operates by transforming the original correlated variables into a new set of orthogonal variables, known as principal components, which are ranked according to the proportion of variance they explain in the dataset [60]. This transformation identifies the most informative directions in the feature space while suppressing low-variance components that are often associated with noise. Given that the FTIR dataset comprised thousands of spectral variables, dimensionality reduction was essential to improve computational efficiency and mitigate the risk of overfitting [60].

The PCA transformation was implemented using the PCA class from the scikit-learn library, with `n_components` set to 0.999. This reduced the original feature space from 3000 spectral variables to 53 principal components while retaining 99.9% of the total variance.

This configuration automatically retained the minimum number of principal components needed to preserve 99.9% of the dataset's total variance, ensuring maximal information retention while reducing feature dimensionality. After fitting the PCA model to the scaled data, the spectra were projected into the reduced principal component space using the `fit_transform` method. The resulting dataset (`data_pca`) provided a compact representation of the original FTIR spectra, where each sample was described by principal component scores rather than raw spectral intensities. The transformed data were subsequently converted into PyTorch tensor representations and organized using DataLoader structures to enable efficient mini-batch training. The overall preprocessing workflow, therefore, consisted of spectral normalization, dimensionality reduction, tensor conversion, dataset partitioning, and model training. Class imbalance was carefully evaluated prior to model development because imbalanced biomedical datasets may introduce classification bias and affect the quality of synthetic data generation. The original dataset contained both diabetic and non-diabetic FTIR spectra in approximately balanced proportions. Consequently, diabetic FTIR spectra were available and included throughout model development. Synthetic data generation was not restricted to the non-diabetic class; instead, synthetic spectra were generated separately for each disease class to preserve class-specific biochemical characteristics and maintain the original class distribution. This strategy avoided introducing artificial class imbalance and ensured that the augmented datasets remained representative of both diabetic and non-diabetic populations. Consequently, no additional balancing techniques, such as random oversampling, random undersampling, SMOTE, or class-weighted loss functions, were required. To preserve class balance throughout the experimental workflow, training, validation, and testing subsets were generated using randomized splitting procedures with a fixed random seed.

To ensure reproducibility and facilitate future deployment, the trained PCA model was serialized and stored using the `joblib` library, enabling the same transformation to be consistently applied to new data without retraining. The dimensionality of the resulting tensor, denoted as `input_dim`, corresponded to the number of retained principal components and served as the input dimension for the generative models. Overall, the PCA preprocessing strategy provided a compact, noise-reduced representation of the FTIR spectra, improving both computational efficiency and model generalization performance.

4.3. Split of the Sample Set

The division of the dataset into distinct subsets is a critical step in the development and evaluation of ML models. In particular, separating data into training and test subsets enables an unbiased assessment of a model's ability to generalize to previously unseen samples. Evaluating a model on the same data used for training would yield overly optimistic performance estimates and fail to reflect real-world applicability [60]. Therefore, an appropriate data splitting strategy is essential to mitigate overfitting and ensure the reliability of the reported results. The dataset comprised 100 FTIR spectra, each represented by 3000 discrete wavenumber points, where each feature corresponded to the absorbance measured at a specific wavenumber within the 400–4000 cm^{-1} spectral range. These features represent individual spectral sampling points rather than aggregated spectral bins, resulting in a data matrix of size 100×3000 . The generative models were trained exclusively using the original FTIR blood serum dataset acquired from original clinical measurements. Following model training, synthetic spectra were generated using the trained AE and GAN models. For evaluation purposes, the resulting synthetic spectra were combined with the original spectra to form a balanced mixed dataset comprising 100 original and 100 synthetic samples (100:100 ratio), yielding a total of 200 samples. This ratio refers solely to the composition of the evaluation dataset and does not imply that synthetic samples were used during the training of the generative models. The balanced dataset ensured

that classification performance was not biased toward a dominant class. To evaluate model performance, the dataset was split into training and test sets at an 80:20 ratio. The train-test split was performed at the sample level, treating each FTIR spectrum with 3000 spectral variables as a single observation. After combining the original and synthetic datasets, a total of 200 samples were available. An 80:20 split was selected because the available dataset was relatively small. This ratio provides a commonly adopted balance between maximizing the number of samples available for model learning (160) and preserving an independent test set (40) for unbiased performance evaluation. More conservative splits, such as 70:30 or 60:40, would further reduce the amount of training data available to both the generative and classification models, potentially increasing estimation variance and reducing model stability. Therefore, the 80:20 ratio was considered an appropriate compromise for this exploratory study. The 160 training samples were further split into training and validation sets at an 80:20 ratio, yielding 128 for training and 32 for validation. The validation subset was used to monitor model convergence, optimize hyperparameters, and assess potential overfitting during training. This partitioning procedure was implemented using the `random_split` function from the PyTorch library, which randomly assigns samples to each subset. To ensure reproducibility of the experimental results, a fixed random seed (random state = 42) was used throughout the splitting process. To further improve the statistical reliability of classifier performance estimates and mitigate limitations associated with the relatively small test set, classifier evaluation was additionally performed using stratified 10-fold cross-validation. Performance metrics were calculated independently for each fold and reported as mean $\pm$ standard deviation across all folds. This approach provides a more robust estimate of model generalization performance and reduces sensitivity to any particular data partition. After dataset partitioning, data loading was performed using the `DataLoader` utility in PyTorch. This component enables efficient mini-batch processing, which is essential for scalable model training. A batch size of 64 samples was selected as a compromise between computational efficiency and memory constraints. The training data loader was configured with shuffling enabled (`shuffle = True`), ensuring that the order of samples was randomized at the beginning of each training epoch. This randomness helps reduce the risk of the model learning spurious correlations associated with sample ordering and improves generalization performance. In contrast, the testing data loader was configured with shuffling disabled (`shuffle = False`), as deterministic ordering is preferred during evaluation to maintain consistency and reproducibility of results. During iteration, each batch returned by the data loader was structured as a tuple, where the first element corresponded to the input feature tensor and subsequent elements (if present) corresponded to labels or auxiliary information. In this implementation, the input data was accessed through `batch[0]`. The dimensionality of each batch reflected both the batch size and the feature space (up to 64×3000), and inspection of these tensor shapes served as an important validation step to confirm correct data handling before model training.

Overall, this data partitioning and loading strategy provides a robust experimental framework by preventing data leakage, enabling fair model evaluation, supporting efficient training, and improving the reliability of reported performance estimates.

4.4. Validation Loss Curve

The validation process is employed to assess model performance on an independent dataset that is not used for parameter updates during training. This provides an estimate of the model's ability to generalize to unseen data and facilitates the detection of potential overfitting. During validation, no gradient updates are performed; instead, the model operates in inference mode. Specifically, all network components are switched to evaluation mode using the `(.eval())` setting in PyTorch, ensuring deterministic behavior of layers

such as dropout and batch normalization. Furthermore, gradient computation is disabled using the (`torch.no_grad()`) context, reducing memory usage and computational overhead. Validation is conducted at the end of each training epoch. The model iterates over the validation dataset in mini batches, and the corresponding loss values are accumulated and averaged across all batches to obtain epoch-level validation metrics. These metrics are recorded to track performance trends over time. For GAN-based models, validation involves evaluating both the generator and discriminator networks. During each validation step, original samples from the validation dataset are first passed to the discriminator to obtain predictions for the original data. In parallel, synthetic samples are generated by feeding latent vectors, sampled from a predefined noise distribution, into the generator. These generated samples are subsequently evaluated by the discriminator to produce predictions for fake data. The discriminator validation loss is computed as the average of two components: (i) the loss between the discriminator's predictions on original samples and their corresponding ground truth labels (set to 1), and (ii) the loss between predictions on generated samples and fake labels (set to 0). This is formally defined as expressed in Equation (3):

$$L_D = \frac{L_{real} + L_{fake}}{2} \quad (3)$$

where L_{real} denotes the loss on real samples and L_{fake} denotes the loss on generated samples. The generator validation loss evaluates the ability of the generator to produce samples that are classified as real by the discriminator. This is computed by comparing discriminator predictions on generated samples against real labels as mentioned in Equation (4):

$$L_G = L_{fake \rightarrow real} \quad (4)$$

For each epoch, the discriminator and generator validation losses are accumulated across all mini batches and averaged. The resulting values are stored (`d_val_losses` and `g_val_losses`) to monitor adversarial training dynamics over time. A stable training process is typically characterized by balanced behavior between the generator and discriminator. In contrast, divergence or oscillatory patterns in validation losses may indicate training instabilities such as mode collapse, vanishing gradients, or overfitting. For AE models, validation focuses on reconstruction performance rather than adversarial learning. During each validation step, input samples are passed through the encoder–decoder architecture to produce reconstructed outputs. The validation loss is computed by comparing the reconstructed outputs with the original inputs using a reconstruction loss function, such as Mean Squared Error (MSE) or Binary Cross Entropy (BCE), depending on the nature of the data. This loss quantifies the model's ability to accurately capture the underlying structure of the input data. As with GAN, the reconstruction loss is accumulated over all mini batches and averaged to obtain the epoch-level validation loss. These values are recorded to monitor convergence and generalization performance. A well-performing AE is typically characterized by a decreasing and stable validation loss. Conversely, an increasing gap between training and validation loss may indicate overfitting, while persistently high validation loss may suggest underfitting or insufficient model capacity.

5. Dataset and Algorithm Setup

This section provides a detailed description of the algorithm's workflow, including its primary steps and sub-processes, and explains their respective roles in generating synthetic data. In addition, it specifies the computational platform utilized for data generation and outlines the configuration of the algorithm parameters. To ensure reproducibility and clarity, the following subsections describe the dataset characteristics, preprocessing pro-

cedures, model training strategy, and synthetic data generation methodology. The original FTIR blood serum dataset used for model training was obtained from a previously published study [61], consisting exclusively of FTIR blood serum spectra collected during clinical measurements, conducted under the ethical approval and informed consent procedures reported therein. The dataset used in the present study contained only anonymized spectral measurements and did not include personally identifiable information. A total of 100 original spectra were available, comprising diabetic and non-diabetic samples as reported in the source dataset [61]. No synthetic data were used during the training of either the AE or GAN models. Synthetic spectra were generated only after model training and then combined with the original spectra for evaluation. Consequently, the mixed evaluation dataset contained 100 original spectra and 100 synthetic spectra (200 samples in total), whereas the generative models themselves were trained exclusively on original clinical measurements. The implementation was conducted in Python 3.13.1 within the Anaconda Navigator environment using Jupyter Notebook 7.2.2. Deep learning model development was supported by the PyTorch framework, while the Matplotlib and Seaborn libraries were employed for data visualization. This computational setup facilitated efficient data preprocessing, algorithm execution, and result interpretation. To prepare the data for model training, the tabular dataset, including biochemical markers relevant to disease prediction, underwent a series of preprocessing and transformation steps. The algorithmic workflow begins with data loading, where the diabetic dataset stored in CSV format is imported and made available for subsequent preprocessing and training. Data normalization follows, scaling all numerical features to a common range (e.g., 0–1 or −1–1) to prevent large, valued variables from dominating model learning and to enhance training stability [62]. Subsequently, PCA is applied to reduce data dimensionality while preserving most of the variance, thereby simplifying the dataset, removing noise, and accelerating computation [63]. The processed data are then converted into PyTorch tensors to ensure compatibility with the deep learning framework [63]. The processed dataset was partitioned according to the strategy described in Section 4.3 before model training and evaluation [64]. A DataLoader is then created to organize the data into mini-batches, manage shuffling operations, and improve training efficiency and randomness [65]. Following data preparation, the model definition phase is initiated. In the AE architecture, the encoder compresses the input data into a lower-dimensional latent representation, while the decoder reconstructs the original data from that latent space [52]. In contrast, the GAN architecture consists of a generator that produces synthetic samples resembling the original data and a discriminator that distinguishes between original and generated samples [58]. After defining the architectures, parameter settings such as learning rate, batch size, number of epochs, and latent space dimension are optimized to guide the learning process and influence the fidelity of the generated data. During the training loop, data are iteratively fed through the model, weights are updated, and the loss function is minimized. This iterative process enables the network to learn complex patterns present in the original dataset and generate synthetic data with similar statistical and structural properties [66]. Multiple training iterations across successive epochs ensure convergence and stable model performance. Generating synthetic datasets using GAN and AE models required specific algorithmic configurations, as shown in Table 1. For the GAN model, the final configuration consisted of 10,000 training epochs, a batch size of 64, a discriminator learning rate of 0.0001, and a generator learning rate of 0.0002. For the AE model, the final configuration included 1000 training epochs, a learning rate of 0.002, and an encoding dimension of 256. The architecture and hyperparameter values were selected through an empirical optimization procedure.

Table 1. Hyperparameter configurations were evaluated during model optimization, and the final selected parameters for the AE and GAN models.

Model	Parameter	Tested Values	Final Value	Selection Criterion
AE	Latent Dimension	64, 128, 256, 512	256	Best reconstruction-quality/complexity trade-off
AE	Learning Rate	0.0005, 0.001, 0.002, 0.005	0.002	Fastest stable convergence
AE	Epochs	500, 1000, 1500	1000	Validation loss plateau
GAN	Generator Learning Rate	0.0001, 0.0002, 0.0005	0.0002	Stable adversarial training
GAN	Discriminator Learning Rate	0.00005, 0.0001, 0.0002	0.0001	Best generator–discriminator balance
GAN	Epochs	5000, 10,000, 15,000	10,000	No significant improvement beyond 10,000 epochs
GAN	Batch Size	32, 64, 128	64	Stable convergence and memory efficiency

The hyperparameter optimization process was conducted using an empirical trial-and-error strategy guided by validation performance. Multiple combinations of learning rates, latent dimensions, batch sizes, and training epochs were evaluated. Model selection was based on four criteria: (i) convergence stability assessed through training and validation loss curves; (ii) reconstruction quality for the AE model; (iii) adversarial training stability for the GAN model; and (iv) the realism and statistical fidelity of the generated spectra. The final configuration was selected as the best compromise between synthetic data quality, training stability, and computational efficiency. The network architectures were intentionally kept relatively simple due to the limited size of the FTIR dataset, thereby reducing the risk of overfitting while preserving the ability to capture the data's principal spectral characteristics. Among the tested latent space dimensions, 256 provided the best tradeoff between reconstruction accuracy and computational complexity. Latent dimensions of 64 and 128 resulted in higher reconstruction losses, indicating insufficient representation capacity, whereas increasing the latent dimension to 512 produced only marginal improvements while substantially increasing computational cost. Similarly, a batch size of 64 provided stable convergence and efficient training, whereas smaller batch sizes increased training variance and larger batch sizes slowed convergence. For the GAN model, a generator learning rate of 0.0002 and a discriminator learning rate of 0.0001 produced the most stable adversarial dynamics and yielded synthetic spectra with the highest spectral similarity to the original data. These configurations were therefore selected to optimize the learning process and ensure reliable model convergence.

Finally, the evaluation phase involved applying ML classifiers, including Support Vector Machines (SVM) and Random Forests (RF), to assess the quality and usability of the synthetic data [62]. A confusion matrix was subsequently used to compare predicted and actual labels, thereby evaluating whether the synthetic data preserved predictive characteristics comparable to those of the original dataset [62].

5.1. Classifiers Evaluation

When generating synthetic datasets, it is essential to evaluate whether the generated data accurately reproduces the statistical characteristics of the original distribution. Because the available dataset was relatively small, classifier performance was evaluated using stratified k-fold cross-validation rather than relying solely on a single train–test split. This approach provides a more robust estimate of model generalization performance and reduces sensitivity to a particular dataset partition. It is important to distinguish between synthetic data realism assessment and disease classification, as these represent two fundamentally different evaluation objectives. In the realism assessment task, classifiers are trained to identify whether a sample originates from the original or synthetic dataset. In this context, poor classifier performance, with accuracy approaching random guessing,

indicates high synthetic-data realism because the classifier is unable to reliably distinguish generated spectra from original spectra. Conversely, in the diabetes-classification task, classifiers are trained to distinguish diabetic from non-diabetic samples, where high accuracy, sensitivity, specificity, F1-score, and AUC values are desirable. Consequently, the interpretation of classifier performance depends on the specific evaluation objective. This study demonstrates the effectiveness of GAN and AE models in generating synthetic FTIR datasets for diabetes-related analysis and for distinguishing between diabetic and non-diabetic samples. One approach for evaluating synthetic data realism is to train classifiers to discriminate between original and synthetic spectra. If the generated data are unrealistic, classifiers will easily identify distinguishing patterns and achieve high classification accuracy. In contrast, if the synthetic data closely resembles the original distribution, classifier performance will approach random guessing, indicating high realism. Therefore, high classifier accuracy suggests that the synthetic data remain distinguishable from the original data, whereas low classifier accuracy, particularly values approaching 50%, indicates that the generated data are highly realistic. This distinction is important because a synthetic dataset may be realistic in terms of distribution while not necessarily improving disease classification performance, and vice versa. To further reduce optimistic bias, synthetic-data realism and diabetes-discrimination analyses were interpreted separately throughout the study. For original-versus-synthetic discrimination, lower accuracy and AUC values are preferable because they indicate that classifiers cannot easily identify the origin of a spectrum. For diabetic versus non-diabetic discrimination, higher accuracy, recall, specificity, F1-score, and AUC are preferable, as they indicate preservation of disease-relevant spectral information. Consequently, the two evaluation tasks assess different aspects of model performance and were interpreted independently throughout the analysis. Multiple classifiers were employed to comprehensively evaluate the generated data, including Support Vector Machines (SVM), Random Forests (RF), K-Nearest Neighbors (KNN), Gradient Boosting (GB), Logistic Regression (LoR), and Decision Trees (DT). The use of multiple classifiers is important because different algorithms capture different relationships within the data, such as linear separability, nonlinear patterns, local neighborhood structures, and ensemble-based decision boundaries. If all classifiers fail to distinguish between original and synthetic spectra, the confidence that the generated data successfully captures the underlying structure of the original dataset increases. DT is a supervised ML algorithm used for both classification and regression tasks. It functions by recursively splitting the dataset into branches based on decision rules derived from feature values, forming a tree-like structure in which internal nodes represent attribute-based decisions and leaf nodes represent predicted outcomes. The interpretability and simplicity of DT models make them suitable for medical decision-support systems, although they may be prone to overfitting when the tree becomes excessively complex [67]. SVM is a robust supervised-learning algorithm primarily used for classification and, less frequently, for regression. SVM identifies the optimal hyperplane that maximizes the margin between classes, where the nearest observations are known as support vectors. This margin maximization improves generalization to previously unseen data. SVMs are particularly effective in high-dimensional feature spaces and are widely used in biomedical signal and image-classification applications [62]. GB is an ensemble-learning technique that constructs a strong predictive model by combining multiple weak learners, typically decision trees. The model is developed sequentially, with each new tree correcting the residual errors of the previous one. By employing gradient-descent optimization to minimize prediction error, GB often achieves high accuracy and strong robustness. Consequently, it has been widely applied in structured data analysis and clinical risk prediction studies [68]. KNN is a non-parametric, instance-based learning algorithm suitable for both classification and regression tasks. It assumes that observations with similar feature values are close to one

another in the feature space. For classification tasks, KNN assigns class labels based on the majority vote of the k nearest neighbors. For regression tasks, outcomes are predicted by averaging neighboring observations. Despite its simplicity, KNN often performs well when data exhibit clear local structures [69,70]. RF is a widely used ensemble learning algorithm that improves predictive accuracy and reduces overfitting by combining multiple decision trees. Each tree is trained on randomly selected subsets of both samples and features through a bagging strategy. For classification tasks, predictions are determined by majority voting, whereas regression outputs are obtained by averaging individual tree predictions. RF is recognized for its robustness and strong generalization performance in biomedical data-analysis applications [71]. LoR is a statistical and ML algorithm primarily designed for binary classification problems, although extensions to multiclass settings are also possible. It estimates class membership probabilities using the logistic (sigmoid) function, which maps real-valued inputs to probabilities between 0 and 1. The model coefficients provide interpretable estimates of the direction and magnitude of relationships between predictors and outcomes, making LoR particularly attractive in clinical prediction studies [72]. To evaluate the synthetic data, classifier performance was assessed using confusion matrix analysis together with standard classification metrics, including accuracy, precision, recall, and F1-score, which are widely used in classification tasks [62]. These metrics are calculated from the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as defined in Equations (5)–(8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1\ score = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

5.2. Experimental Analysis Method

After being applied to the provided dataset, a preliminary experimental validation of the method was implemented through the FTIR spectra of whole blood samples, locally collected from healthy and diabetic donors. The blood sample collection followed the Ethical approval CEICVS 039/2023 from the Ethics Commission for Research in Life and Health Sciences (CEICVS), from the University of Minho, Portugal. FTIR spectra were acquired from whole blood samples collected from 7 adult volunteers (mean age: 34.4 years), including 6 healthy individuals (2 men and 4 women) and 1 diabetic male volunteer, all with informed consent. Besides gender, age, and disease status (healthy/pathological), no additional information was collected from the volunteers. Statistical analyses were performed exclusively on the 6 healthy samples to characterize variability among normal spectra. All blood samples were acquired on the same day of the measurements, collected into tubes with ethylenediaminetetraacetic acid (EDTA) to prevent blood coagulation, and kept refrigerated until spectra acquisition, avoiding sample deterioration and oxidation. All FTIR measurements were recorded over the spectral range 400–4000 cm^{-1} at a resolution of 4 cm^{-1} , with each spectrum representing the average of 16 scans, using an FTIR spectrometer (PerkinElmer Precisely, Spectrum 100; Shelton, CT, USA). For the spectra acquisition of the fluidic blood samples, a mechanical metallic adapter was used, fitting

in the IR spectrometer sample holder. The adapter includes two CaF₂ windows able for infrared transmission, separated by mylar spacers with 25 µm thickness. For each assay, 15 µL of the blood sample was pipetted into the spacer open area, followed by closing the CaF₂ window and promptly analysing. Each assay was repeated 3 times. Between each measurement, the adapter, mylar spacer, and CaF₂ windows were carefully cleaned with diluted Isopropyl Alcohol (IPA). The absorbance of each sample was calculated using the standard relation in Equation (9).

$$A = 2 - \log_{10}(T(\%)) \quad (9)$$

where (*A*) represents absorbance and (*T*) transmission.

Because only one type-2 diabetic volunteer was available for this preliminary whole blood experiment, the resulting dataset is insufficient for meaningful statistical evaluation of diabetes classification performance. Consequently, the experimental analysis presented in this section should be interpreted exclusively as a proof-of-concept assessment of spectral feasibility and methodological transferability rather than as evidence of diagnostic accuracy, sensitivity, specificity, or clinical classification performance.

6. Results and Discussion

This section presents the experimental results obtained from the proposed synthetic data generation framework and discusses their implications for FTIR diabetes analysis. To provide a comprehensive evaluation, the results are organized into three complementary perspectives. First, the realism of the synthetic spectra generated by the AE and GAN models is assessed through classifier discrimination, spectral similarity analysis, and statistical fidelity metrics. Second, the utility of the generated data for diabetes discrimination is evaluated using multiple ML classifiers and standard performance metrics. Finally, a preliminary experimental validation using independently acquired whole-blood FTIR spectra is presented to assess spectral feasibility and methodological transferability in a more application-oriented setting. Because these analyses address different objectives, their results are interpreted independently throughout this section.

6.1. Synthetic Data Realism Assessment

This subsection evaluates synthetic data realism by assessing the ability of ML classifiers to distinguish original FTIR spectra from synthetic spectra generated by the AE and GAN models. Unlike diabetes classification, where higher classification performance is desirable, the objective of realism assessment is to determine how closely the synthetic data resemble the original data distribution. Consequently, lower classification accuracy, lower AUC values, and reduced separability between original and synthetic samples indicate greater realism of synthetic data and stronger distributional similarity. Before presenting the experimental results, it is important to summarize the dataset structure used throughout this study. The original dataset consisted of FTIR spectra of blood serum collected from diabetic and non-diabetic subjects. Each spectrum was represented by 3000 spectral variables corresponding to absorbance measurements acquired across the FTIR spectral range. Table 2 summarizes the datasets used in each stage. These spectral variables served as the input features for the ML models, while the target variable was the class label indicating diabetic or non-diabetic status. The resulting principal component scores were used as inputs to the AE and GAN models for generating synthetic data. The generated synthetic samples were subsequently combined with the original data and evaluated using multiple ML classifiers. All classification analyses reported in the following sections were performed using these processed feature representations and their associated class labels.

Table 2. Distribution of original clinical spectra and synthetic spectra across the different stages of model training and evaluation.

Dataset Stage	Original Samples	Synthetic Samples	Total
Generative model training	100	0	100
AE-generated dataset	100	100	200
GAN-generated dataset	100	100	200
Classifier evaluation	100	100	200

6.1.1. Loss Function

The GAN and the AE models were trained independently to generate synthetic tabular datasets. Model weights were optimized through backpropagation, with convergence in the loss functions indicating successful learning. The training and validation loss curves for the GAN and AE are presented in Figures 4a and 4b, respectively, both exhibiting high initial losses that gradually stabilize over epochs. Upon completion of training, the models were used to generate synthetic data, whose fidelity was first assessed by comparing the spectra of the original and synthetic datasets. The GAN loss plot provides valuable insight into the interaction between the generator and discriminator during training. Figure 4a displays four curves across epochs. The blue curve shows the discriminator training loss, indicating how effectively the discriminator distinguishes between original and synthetic samples. The orange curve corresponds to the generator training loss, reflecting the generator's ability to produce realistic samples that can fool the discriminator. The green-dashed curve denotes the discriminator's validation loss, which measures its generalization to unseen data. The red dashed curve represents the generator validation loss, indicating how realistic the generated samples appear to the discriminator on the validation set. Loss values fluctuating between 0.6 and 0.8 suggest a balanced GAN, where neither model dominates the other. The discriminator loss (0.6–0.7), which remains close to the generator loss (0.7–0.8), indicates stable adversarial dynamics. These oscillations are expected due to the simultaneous optimization of two competing networks. Diverging loss trends, by contrast, would imply instability or mode collapse. In this case, both training and validation losses oscillate around similar means, demonstrating a stable equilibrium and good generalization, with no evidence of overfitting. Similarly, regarding the AE training and validation loss curves, Figure 4b shows the model convergence behavior. During the initial phase (0–100 epochs), both losses decrease sharply as the model rapidly learns to reconstruct input data. In the middle phase (100–400 epochs), the losses continue to decrease more gradually, converging closely, indicating strong generalization without overfitting. In the later phase (400–1000 epochs), both curves plateau, with the training loss slightly lower than the validation loss. This small and consistent gap suggests that the AE has successfully converged, achieving low reconstruction error and good generalization with average validation loss (0.00158). The model was evaluated over 10 consecutive intervals of 100 epochs each, spanning the training range from 0 to 1000 epochs. Extending training beyond approximately 400 epochs yields negligible performance improvement, confirming that the model has reached its optimal learning state.

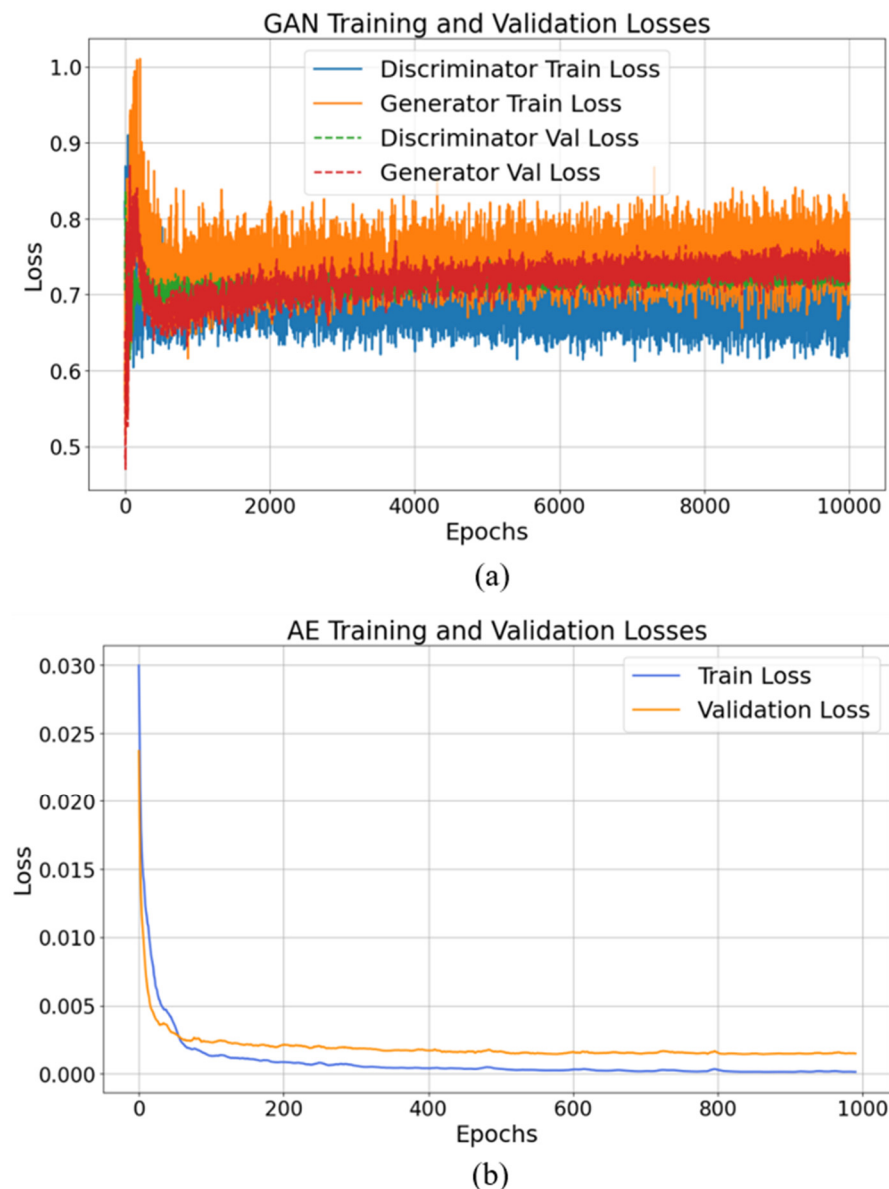


Figure 4. Loss function plots during the training for both models: (a) GAN; (b) AE.

The loss function trends demonstrate stable convergence across epochs, confirming efficient training dynamics for both models. The GAN achieved balanced adversarial training between its generator and discriminator components, while the AE exhibited minimal overfitting and robust generalization, signifying a healthy reconstruction process.

6.1.2. Performance Metrics for Original Versus the Synthetic Blood Spectra

Classifier performance for the GAN-generated and AE-generated datasets is summarized in Tables 3–5. The evaluation was performed using separate training and testing procedures, along with stratified 10-fold cross-validation, to assess classifier generalization and the distinguishability of synthetic spectra from original spectra. Because the objective of this experiment was synthetic data realism assessment rather than disease classification, classifier performance should be interpreted differently from conventional classification tasks. In this context, lower classification accuracy indicates that synthetic spectra more closely resemble the original spectra and are therefore more difficult to distinguish. Table 3 presents the results for the testing set. For the GAN-generated dataset, test-

ing accuracies ranged from 31% to 59%, with DT achieving the highest accuracy (59%) and LoR the lowest (31%). Precision, recall, and F1-score values were similarly modest across most classifiers. These results indicate that the classifier's ability to distinguish original spectra from GAN-generated spectra is limited. In contrast, the AE-generated dataset produced substantially higher testing accuracies, ranging from 59% to 81%, with GB achieving the highest accuracy (81%), followed by SVM (78%) and DT (73%). The consistently higher precision, recall, and F1-score values obtained for AE-generated data indicate that classifiers could more readily identify differences between original and synthetic spectra. Table 4 summarizes the training set performance. For the GAN-generated dataset, training accuracies ranged from approximately 58% to 85%, whereas the AE-generated dataset produced accuracies ranging from approximately 51% to 91%. In both cases, training performance exceeded testing performance, which is expected because classifiers are optimized on the training data. However, the moderate difference between training and test accuracies suggests that severe overfitting was largely avoided. The largest discrepancies were observed for RF and GB, indicating that these models may have captured additional dataset-specific characteristics, particularly for the AE-generated spectra. To further assess robustness, Table 5 reports results from stratified 10-fold cross-validation. For the GAN-generated dataset, mean accuracies ranged from approximately 49.5% to 61.5%, with LoR achieving the highest average accuracy (61.5%). For the AE-generated dataset, mean accuracies ranged from approximately 51.0% to 84.0%, with GB achieving the highest average accuracy (84.0%), followed by DT (73.0%) and RF (70.5%). The relatively small standard deviations observed across folds indicate reasonably stable classifier performance and support the reliability of the reported results. Taken together, the results demonstrate that classifiers generally found it more difficult to distinguish GAN-generated spectra from original spectra than AE-generated spectra. From a synthetic data realism perspective, this behavior is desirable because it suggests greater distributional similarity between the GAN-generated and original FTIR spectra. Conversely, the higher classification accuracies, precision, recall, and F1-scores obtained for the AE-generated dataset indicate greater separability between synthetic and original samples. This does not imply superior synthetic data quality; rather, it suggests that the AE-generated spectra retained detectable structural characteristics, enabling classifiers to distinguish them more readily from the original data. These findings further support the interpretation that the GAN and AE models emphasize different aspects of data generation. The GAN appears to better reproduce the overall statistical distribution of the original FTIR spectra, thereby reducing classifier separability and increasing perceived realism. In contrast, the AE preserves dominant spectral patterns and discriminative information more strongly, resulting in synthetic spectra that remain more distinguishable from the original dataset. Consequently, synthetic data realism and downstream classification utility should be regarded as complementary but distinct evaluation objectives when assessing the performance of generative models.

Table 3. Test-set classification performance metrics for distinguishing original and synthetic FTIR spectra in the mixed dataset.

Classifier	GAN				Classifier	AE			
	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
DT	59.00	69.00	62.00	65.00	GB	81.00	89.00	73.00	80.00
RF	50.00	61.29	48.72	54.29	SVM	78.00	78.00	78.00	78.00
GB	50.00	61.00	51.00	56.00	DT	73.00	75.00	73.00	74.00
KNN	44.00	56.00	36.00	44.00	RF	69.00	78.00	55.00	64.00
SVM	43.75	37.50	18.75	25.00	KNN	66.00	92.00	34.00	50.00

LoR	31.00	22.00	5.00	8.00	LoR	59.00	59.00	62.00	61.00
-----	-------	-------	------	------	-----	-------	-------	-------	-------

Table 4. Training-set classification performance metrics for distinguishing original and synthetic FTIR spectra in the mixed dataset.

GAN					AE				
Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
RF	85.00	79.79	93.75	86.21	GB	91.43	100.00	82.86	90.62
DT	71.88	71.60	72.50	72.05	RF	89.29	98.25	80.00	88.19
LoR	71.25	69.77	75.00	72.29	DT	73.00	75.00	73.00	74.00
GB	67.14	62.50	85.71	72.29	KNN	72.86	80.77	60.00	68.85
SVM	66.87	60.98	93.75	73.89	LoR	59.00	59.00	62.00	61.00
KNN	58.13	55.75	78.75	65.28	SVM	51.43	52.50	30.00	38.1

Table 5. Stratified 10-fold cross-validation performance metrics for original versus synthetic FTIR spectra discrimination.

GAN					AE				
Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LoR	61.50 ± 0.05	60.04 ± 0.05	68.00 ± 0.12	63.41 ± 0.07	GB	84.00 ± 0.07	89.21 ± 0.08	78.00 ± 0.04	82.40 ± 0.09
DT	59.50 ± 0.11	58.34 ± 0.12	63.00 ± 0.16	60.08 ± 0.14	DT	73.00 ± 0.12	75.00 ± 0.06	73.00 ± 0.10	74.00 ± 0.10
RF	57.00 ± 0.09	55.36 ± 0.08	68.00 ± 0.15	60.62 ± 0.11	RF	70.50 ± 0.08	77.897 ± 0.11	60.00 ± 0.12	66.59 ± 0.10
GB	55.50 ± 0.08	55.50 ± 0.08	61.00 ± 0.04	58.04 ± 0.06	LoR	69.00 ± 0.16	59.00 ± 0.12	72.00 ± 0.12	61.00 ± 0.10
SVM	54.50 ± 0.07	52.86 ± 0.05	77.00 ± 0.14	62.38 ± 0.07	KNN	68.00 ± 0.05	83.98 ± 0.12	46.00 ± 0.11	58.48 ± 0.08
KNN	49.50 ± 0.12	50.79 ± 0.10	68.00 ± 0.13	57.43 ± 0.09	SVM	51.00 ± 0.03	10.00 ± 0.30	02.00 ± 0.06	03.33 ± 0.01

6.1.3. Classifiers Accuracy Analysis

Figure 5 compares the classification accuracies achieved by six ML classifiers when distinguishing original FTIR blood serum spectra from synthetic spectra generated by the GAN and AE models. The results show that AE-generated data generally achieved higher classification accuracies than GAN-generated data across most classifiers, except for SVM, which achieved slightly higher accuracy with GAN-generated data (55%) than with AE-generated data (51%). For the GAN-generated dataset, classification accuracies ranged from 50% to 62%, with LoR achieving the highest accuracy (62%), followed by DT (60%) and RF (57%). These values are relatively close to chance-level classification, indicating that the classifiers experienced difficulty distinguishing GAN-generated spectra from original spectra. In contrast, the AE-generated dataset produced substantially higher accuracies for most classifiers, ranging from 51% to 84%. The highest accuracy was achieved by GB (84%), followed by DT (73%), RF (71%), LoR (69%), and KNN (68%). These results indicate that the AE-generated spectra were generally more distinguishable from the original spectra than the GAN-generated spectra. However, these findings should be interpreted within the context of synthetic data realism assessment. Unlike conventional classification tasks, higher classification accuracy does not necessarily indicate superior synthetic data quality. Instead, higher accuracy reflects greater separability between the original and synthetic datasets, suggesting that classifiers can more readily identify differences between them. Consequently, the higher accuracies observed for the AE-generated data imply that the synthetic spectra retained detectable structural characteristics that distinguish them from the original spectra. Conversely, the lower accuracies observed for the GAN-generated data suggest greater distributional similarity to the original FTIR spectra,

making synthetic and original samples harder to distinguish. This behavior is consistent with the underlying learning mechanisms of the two generative models. The AE is optimized to reconstruct input spectra and therefore tends to preserve dominant structural patterns present in the original dataset. As a result, reconstructed spectra may retain identifiable characteristics that facilitate discrimination from original samples. In contrast, GANs are trained through adversarial learning to reproduce the overall data distribution and generate samples that are difficult to distinguish from original observations. Consequently, the lower classifier accuracies obtained for the GAN-generated data may indicate greater synthetic data realism rather than inferior model performance. These results highlight that classifier separability and synthetic data realism represent complementary but distinct evaluation criteria and should be interpreted accordingly.

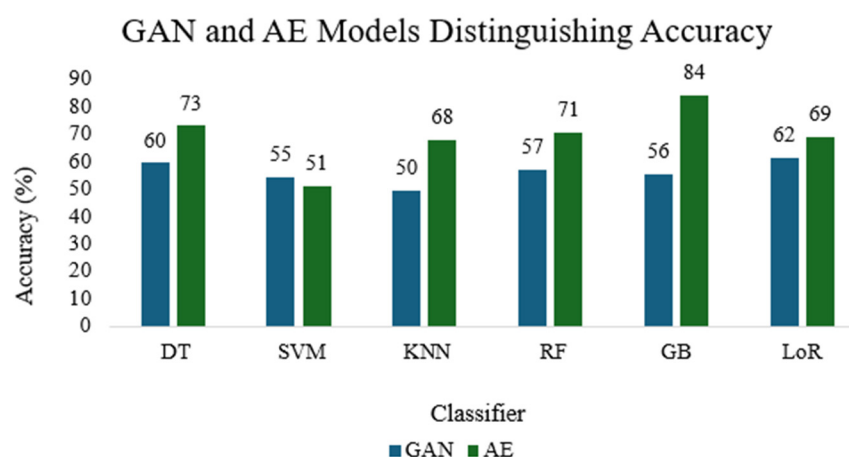
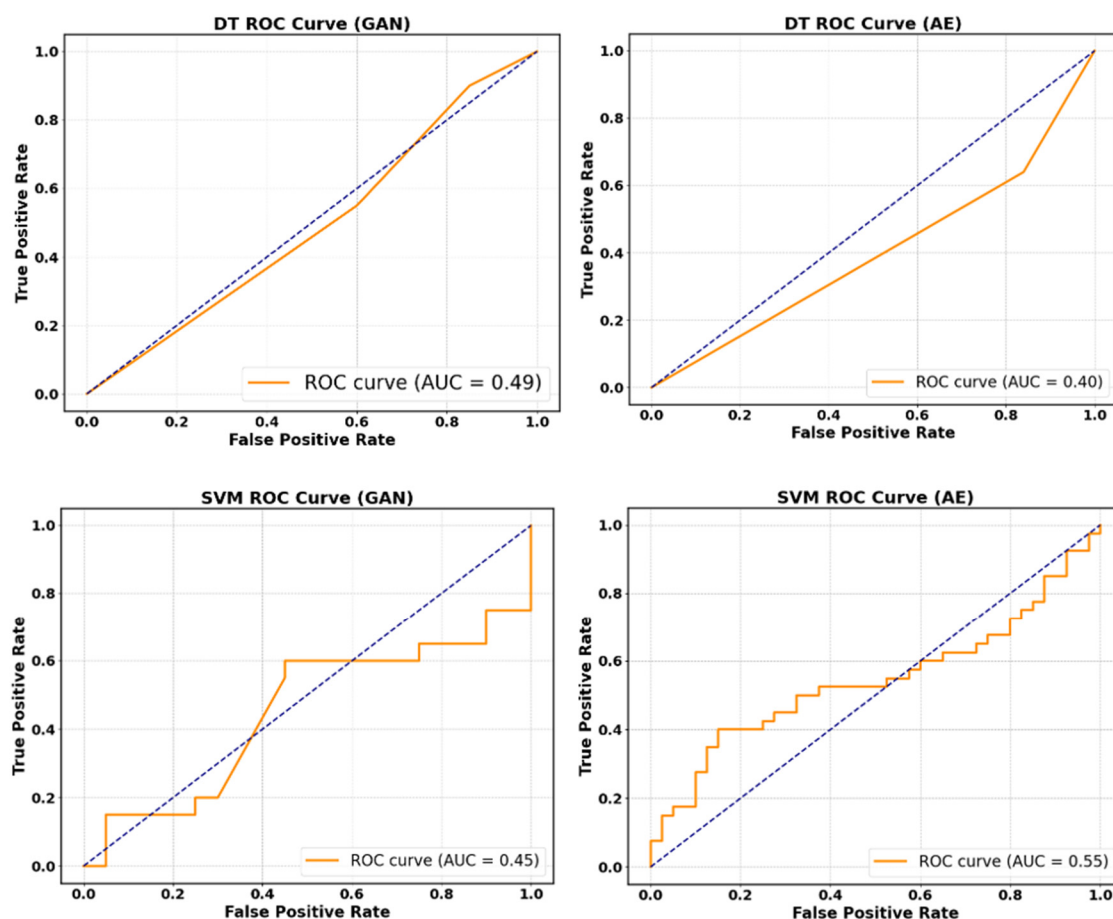


Figure 5. Classification accuracy of the 6 classifier models (DT, SVM, KNN, RF, GB, LoR), measuring the distinguishing efficiency between the GAN and AE models. The blue bar represents the accuracy of the GAN model, and the green bar the accuracy of the AE model.

6.1.4. Receiver Operating Characteristic

A comparative analysis was conducted to evaluate the ability of different ML classifiers to distinguish original FTIR blood serum spectra from synthetic spectra generated by the GAN and AE models. ROC curves and AUC values were used to quantify classifier separability across different decision thresholds. Figure 6 presents the ROC curves obtained for the six evaluated classifiers (DT, SVM, KNN, RF, GB, and LoR). The ROC curves were generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) over multiple classification thresholds, while the diagonal dashed line represents the performance of a random classifier (AUC = 0.5). Overall, the ROC analysis indicates limited classifier separability between the original and synthetic spectra for both generative models. For the GAN-generated dataset, AUC values ranged from approximately 0.45 to 0.61, with GB achieving the highest AUC (0.61), followed by RF (0.58) and LoR (0.55). DT and KNN produced AUC values close to 0.49, while SVM achieved an AUC of 0.45. For the AE-generated dataset, AUC values ranged from approximately 0.40 to 0.55, with SVM achieving the highest AUC (0.55), followed by GB (0.54). RF and LoR produced AUC values of approximately 0.46, KNN achieved 0.44, and DT produced the lowest AUC value of 0.40. The majority of ROC curves for both generative models remain close to the diagonal reference line, indicating that classifier performance approaches chance-level classification. In the context of synthetic data realism assessment, this behavior is desirable because it suggests that the synthetic spectra are highly similar to the original FTIR spectra. Higher AUC values indicate greater separability between original and synthetic sam-

ples, whereas AUC values approaching 0.5 indicate that classifiers are unable to reliably distinguish between the two distributions. Although some variability is observed among classifiers, the overall results suggest that both GAN and AE models generated spectra that were difficult to differentiate from the original data. The slightly higher AUC values observed for GB and RF in the GAN-generated dataset indicate that these classifiers detected limited structural differences between the synthetic and original spectra. Similarly, the slightly higher AUC values obtained by SVM and GB on the AE-generated dataset suggest subtle classifier-specific distinctions. However, because all AUC values remained relatively close to the random classification baseline, the results indicate substantial overlap between the original and synthetic spectral distributions. These findings demonstrate that the synthetic data realism assessment is highly dependent on the classifier employed and that no single classifier consistently provided strong discrimination between original and synthetic spectra. Consequently, the ROC analysis supports the conclusion that both generative models were capable of reproducing the principal characteristics of the original FTIR spectra, while also highlighting the importance of combining ROC analysis with complementary metrics, including accuracy, correlation analysis, and spectral similarity assessment, to obtain a comprehensive evaluation of synthetic data quality and realism.



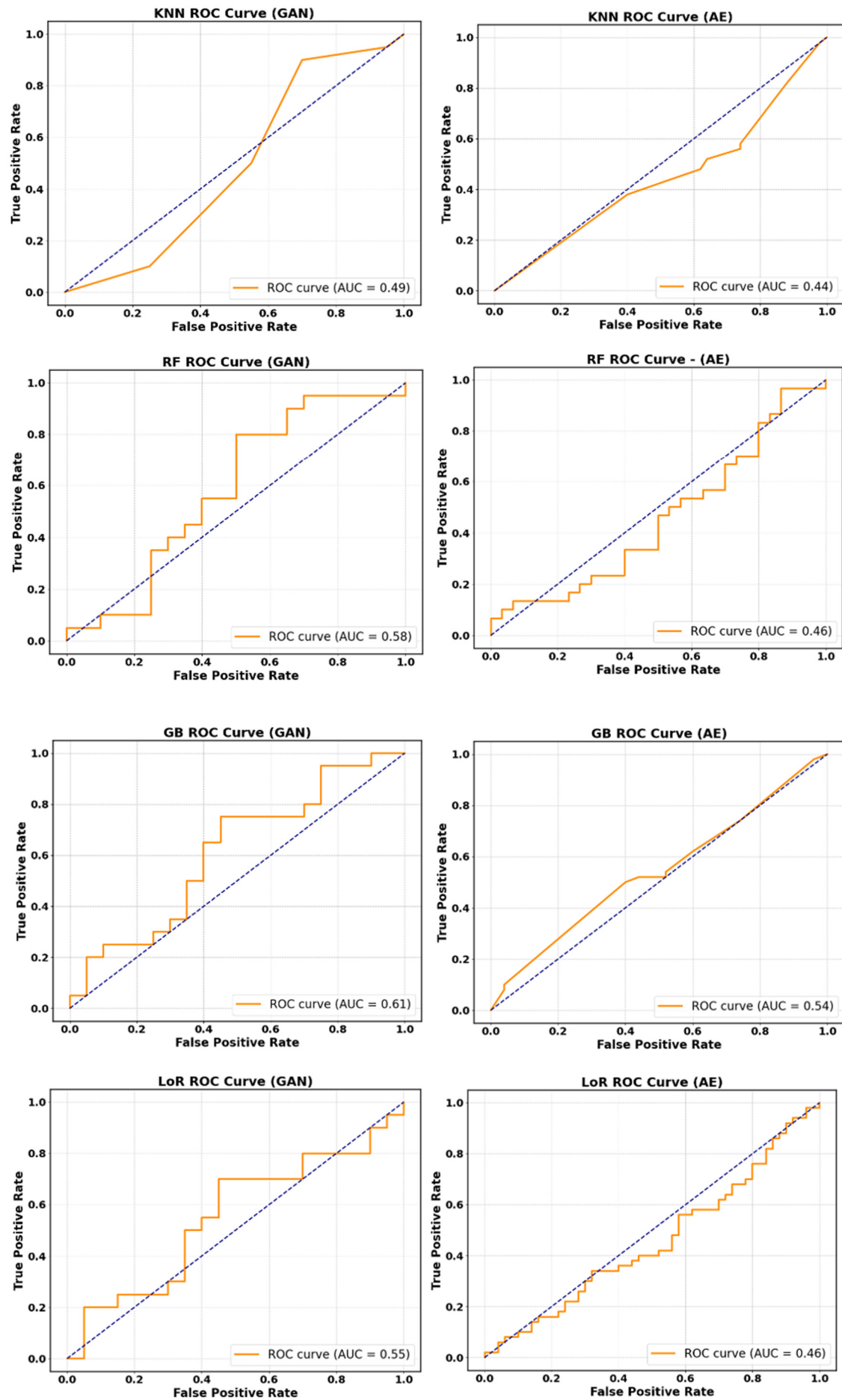


Figure 6. ROC curves for the 6 classifiers across a range of thresholds in both GAN and AE models. The blue dashed diagonal line represents the performance of the classifiers and serves as the baseline reference.

6.1.5. Area Under the Curve for Original Versus Synthetic Spectra

The AUC provides a quantitative measure of classifier separability between original and synthetic spectra across all classification thresholds. An AUC value of 1.0 indicates perfect discrimination, whereas an AUC value of 0.5 corresponds to random classification with no discriminative ability. Figure 7 summarizes the AUC values obtained for the six evaluated classifiers using synthetic spectra generated by the GAN and AE models. Overall, the AUC values for both generative models remained relatively close to the random classification baseline, ranging from 0.45 to 0.61 for the GAN-generated dataset and from 0.40 to 0.55 for the AE-generated dataset. For the GAN-generated spectra, the highest AUC value was achieved by GB (0.61), followed by RF (0.58) and LoR (0.55), whereas SVM produced the lowest AUC value (0.45). For the AE-generated spectra, SVM achieved the highest AUC (0.55), followed by GB (0.54), while DT achieved the lowest AUC (0.40). In most cases, the GAN-generated spectra produced slightly higher AUC values than the AE-generated spectra, although the differences remained modest. From the perspective of synthetic data realism assessment, AUC values close to 0.5 are desirable because they indicate that classifiers are unable to reliably distinguish synthetic spectra from original spectra. Consequently, the generally low AUC values observed for both generative models suggest substantial overlap between the synthetic and original data distributions. The slightly elevated AUC values obtained by GB and RF for the GAN-generated dataset indicate that these classifiers detected only limited structural differences between the synthetic and original spectra. Similarly, the marginally higher AUC value achieved by SVM for the AE-generated dataset suggests that some classifier-specific distinctions remained detectable. Importantly, higher AUC values in this experiment should not be interpreted as evidence of superior synthetic data quality. Instead, higher AUC values indicate greater separability between original and synthetic samples, whereas lower AUC values indicate stronger distributional similarity. Therefore, the overall proximity of the AUC values to the random classification baseline supports the conclusion that both GAN and AE models generated spectra that reproduced the principal characteristics of the original FTIR dataset. These findings further demonstrate that synthetic data realism is classifier-dependent and should be assessed using multiple complementary evaluation metrics rather than relying solely on a single classifier or AUC value.

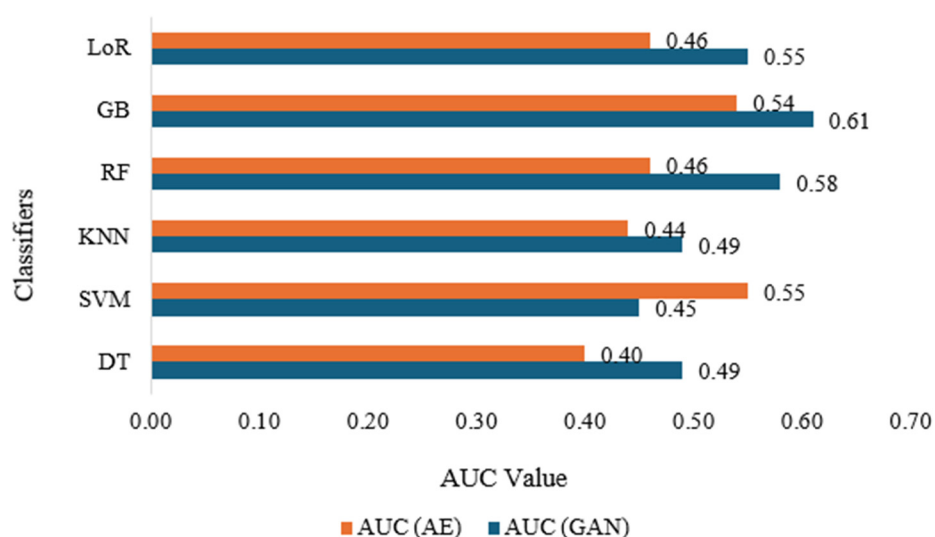


Figure 7. Comparisons between the AUC values of the original and the synthetic for the GAN and the AE models.

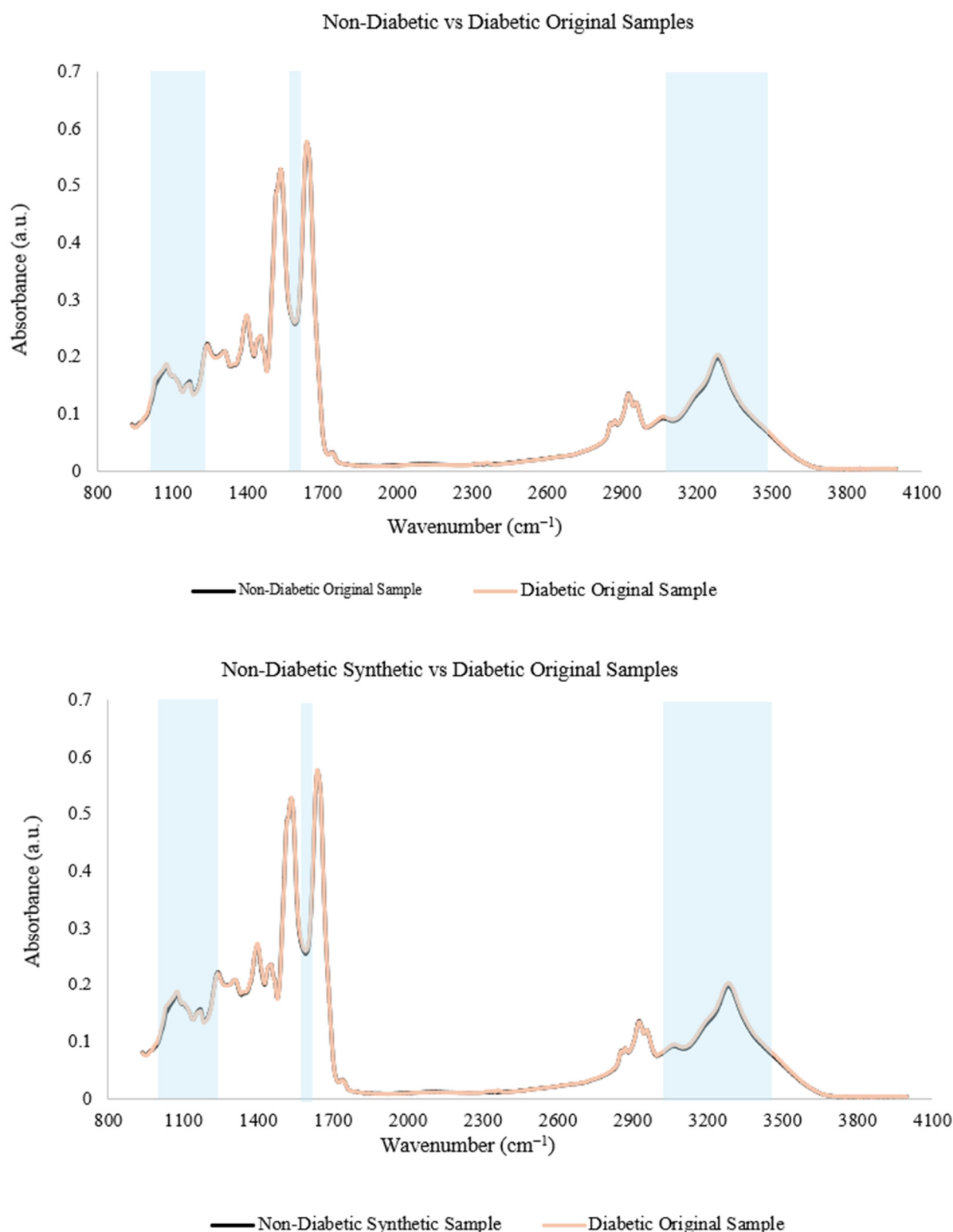
The observed differences in AUC values between the AE and GAN-generated datasets should be interpreted with caution. In the realism assessment task, classifiers were trained to distinguish original spectra from synthetic spectra rather than to classify diabetic and non-diabetic samples. Consequently, a lower AUC does not necessarily indicate lower synthetic data quality. Instead, reduced classifier separability may reflect a closer approximation of the original data distribution, making synthetic samples more difficult to distinguish from original samples. From this perspective, the lower AUC values observed for GAN-generated data may be interpreted as evidence of increased distributional similarity and enhanced realism. Conversely, the higher AUC values obtained for AE-generated data suggest that the synthetic spectra retained more distinguishable characteristics relative to the original dataset. Therefore, the AE and GAN models should not be viewed as exhibiting superior or inferior performance solely based on AUC values. Rather, they represent different trade-offs between preserving discriminative structure and maximizing distributional realism.

6.1.6. Spectral Fidelity Analysis

A visual comparison of the spectral similarity between non-diabetic and diabetic samples, including both the original datasets and the synthetic data generated using GAN and AE, is presented in Figures 8 and 9, respectively.

Each subplot presents a comparison of FTIR spectra from non-diabetic and diabetic samples across various combinations of original and synthetic datasets, including those generated by a GAN and reconstructed using an AE. In all cases, the highlighted light blue regions indicate key wavelengths where subtle differences or peak shifts between the two classes are most pronounced. As illustrated in Figure 8, the synthetic spectra generated by the GAN closely replicate the corresponding original spectra across the entire mid-infrared range. The substantial overlap between the non-diabetic (black) and diabetic (orange) curves demonstrates that the GAN effectively captures the essential biochemical signatures of both classes. Minor variations in absorbance are observed within the highlighted light blue bar fingerprint region ($1000\text{--}1200\text{ cm}^{-1}$), along with a slight peak shift around 1650 cm^{-1} and differences in the ($3000\text{--}3500\text{ cm}^{-1}$) region [73]. These variations indicate subtle, class-specific spectral differences that may reflect underlying biochemical or structural alterations associated with diabetes. Similarly, Figure 9 demonstrates that the AE generates synthetic spectra that are highly consistent with the original data. The near-complete overlap between reconstructed and original spectra suggests that the AE effectively preserves the spectral characteristics of both sample types. The slightly smoother appearance (as evidenced by the uniformity of the light-blue-highlighted regions) of the AE-generated curves indicates that the model produces more stable, less noisy representations than the GAN, while still retaining diagnostically relevant features. In both models, the highlighted regions confirm that, although subtle, biologically meaningful differences between diabetic and non-diabetic samples are preserved in the synthetic datasets. This suggests that both approaches successfully reproduce the underlying data distribution and can be reliably employed for data augmentation or downstream analytical tasks when original data are limited. Furthermore, the ($1000\text{--}1200\text{ cm}^{-1}$) fingerprint region was systematically analyzed in both models (GAN and AE) to identify specific spectral locations where differences between classes occur [74]. This analysis facilitates differentiation between non-diabetic and diabetic samples. Specifically, the spectral profiles of the two classes exhibit a high degree of similarity; therefore, discrimination relies on subtle variations within the fingerprint region. Consequently, individual peak assignments (in the

analytical region) are not emphasized in this section, as they do not provide sufficiently distinct analytical information. Figure 10 illustrates the regions (1000–1200 cm^{-1} and 3000–3500 cm^{-1}) where the peak differences between the non-diabetic and diabetic spectra are in terms of magnitude shifting, whereas Figure 11 shows a slight peak shift difference between the non-diabetic and diabetic spectra [75].



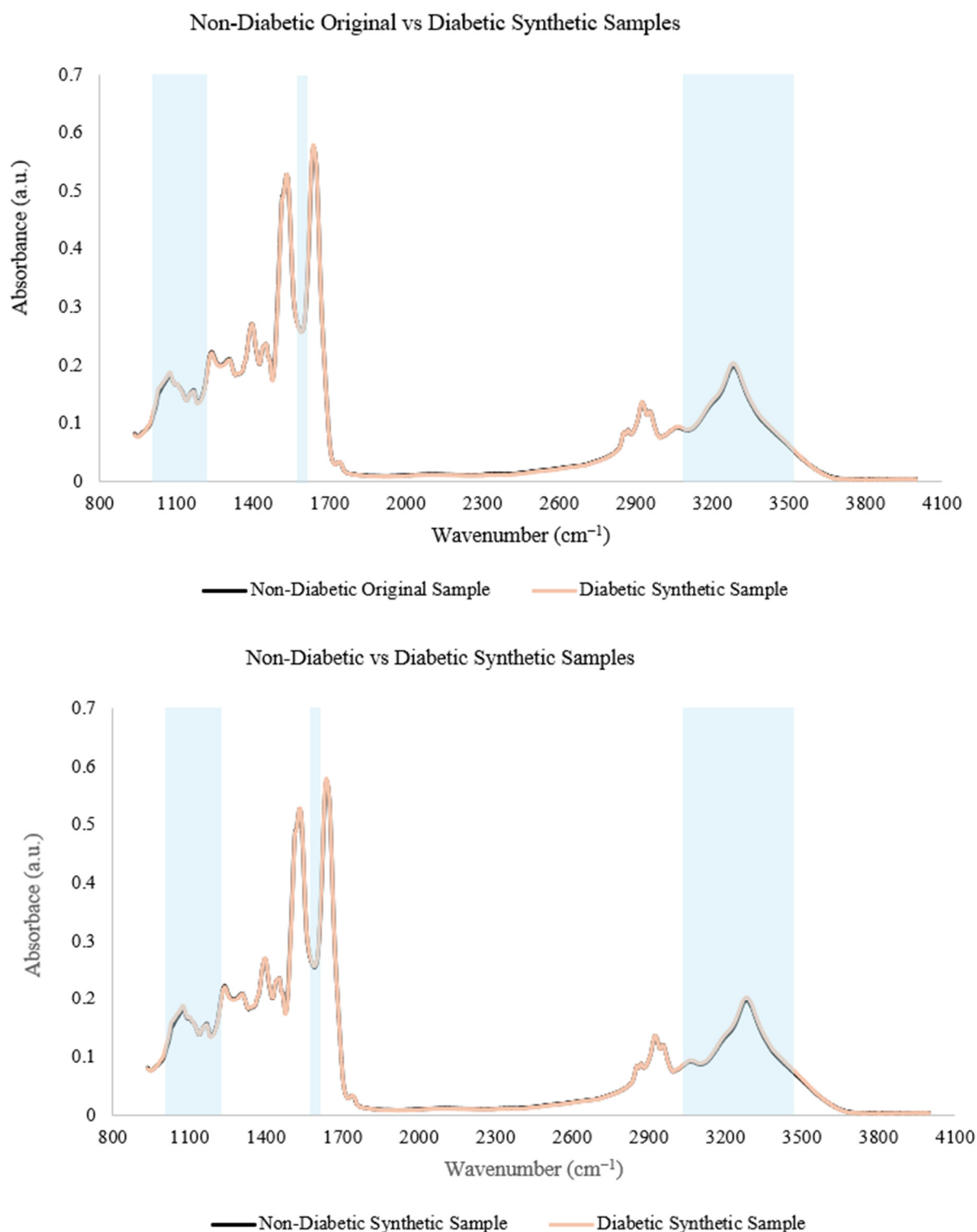
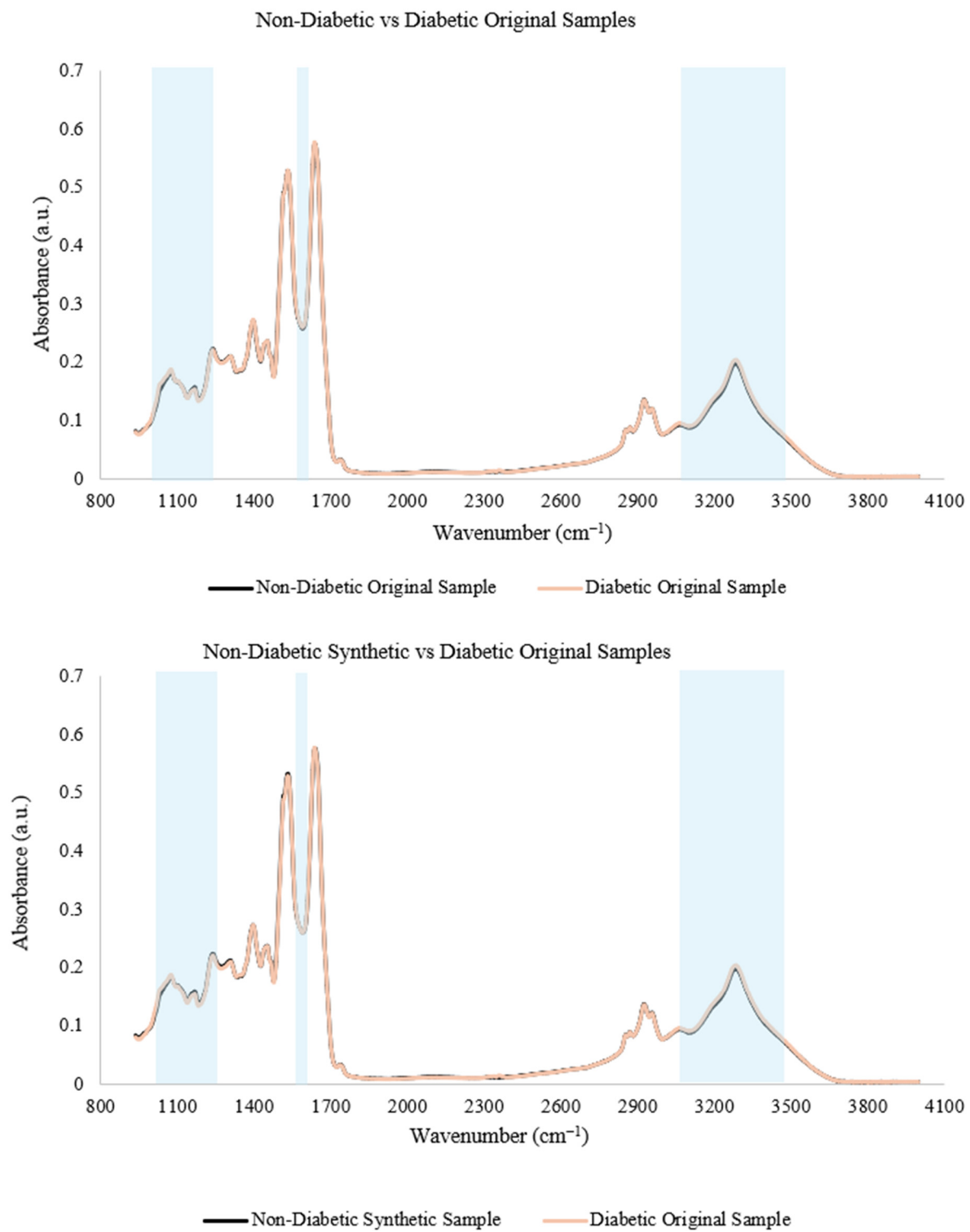


Figure 8. Comparison between the original and synthetic FTIR absorbance spectra, for two blood samples, one diabetic and one non-diabetic, displayed across the full mid infrared wavenumber range (800–4000 cm^{-1}) generated by the GAN model. In each plot, the black spectrum represents the non-diabetic sample, and the orange spectrum represents the diabetic sample. The light blue regions indicate key wavelengths where subtle differences or peak shifts between the two classes are most pronounced.



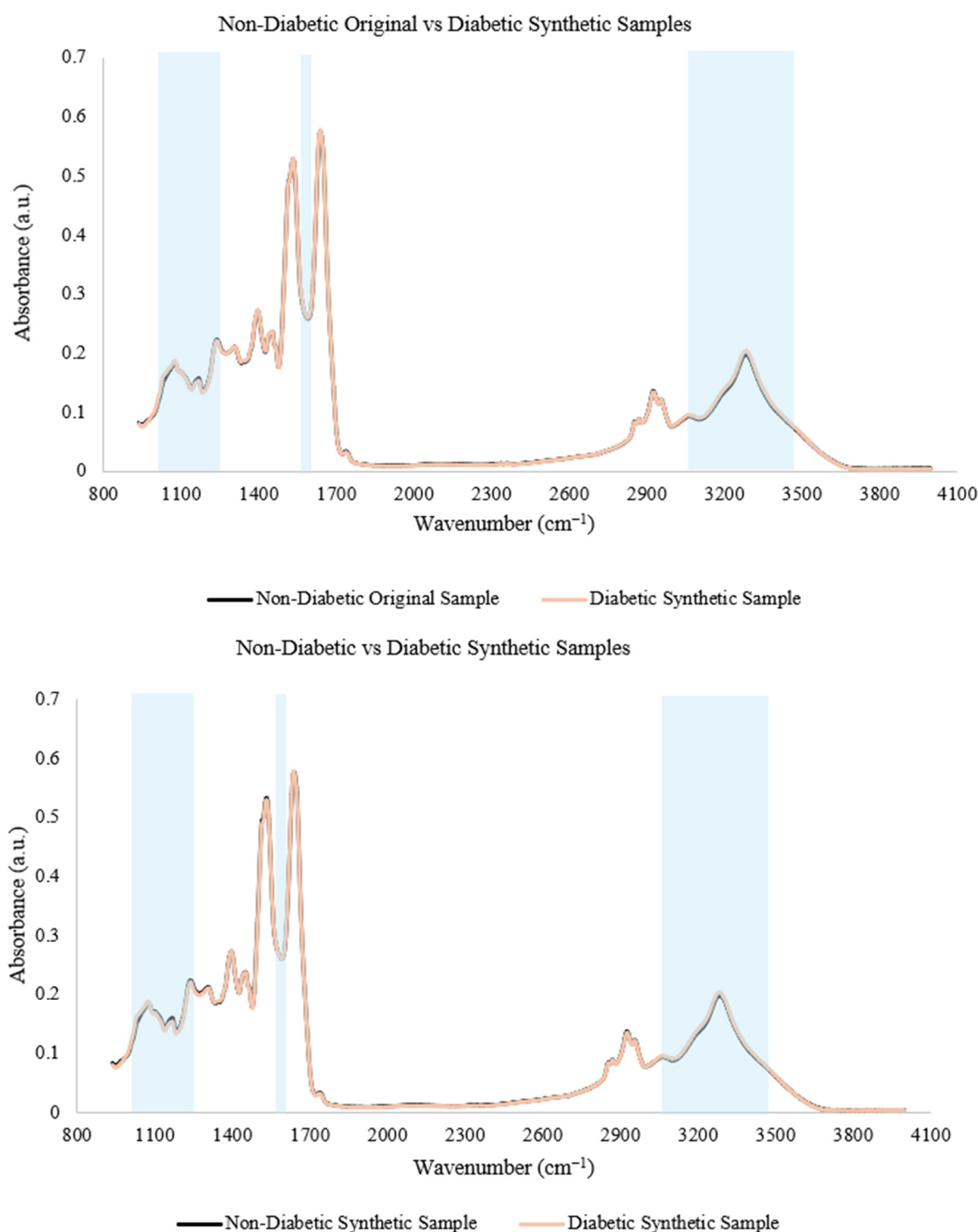


Figure 9. Comparison between the original and synthetic FTIR absorbance spectra, for two blood samples, one diabetic and one non-diabetic, displayed across the full mid infrared wavenumber range (800–4000 cm^{-1}) generated by the AE. In each plot, the black spectrum represents the non-diabetic sample, and the orange spectrum represents the diabetic sample. The light blue regions indicate key wavelengths where subtle differences or peak shifts between the two classes are most pronounced.

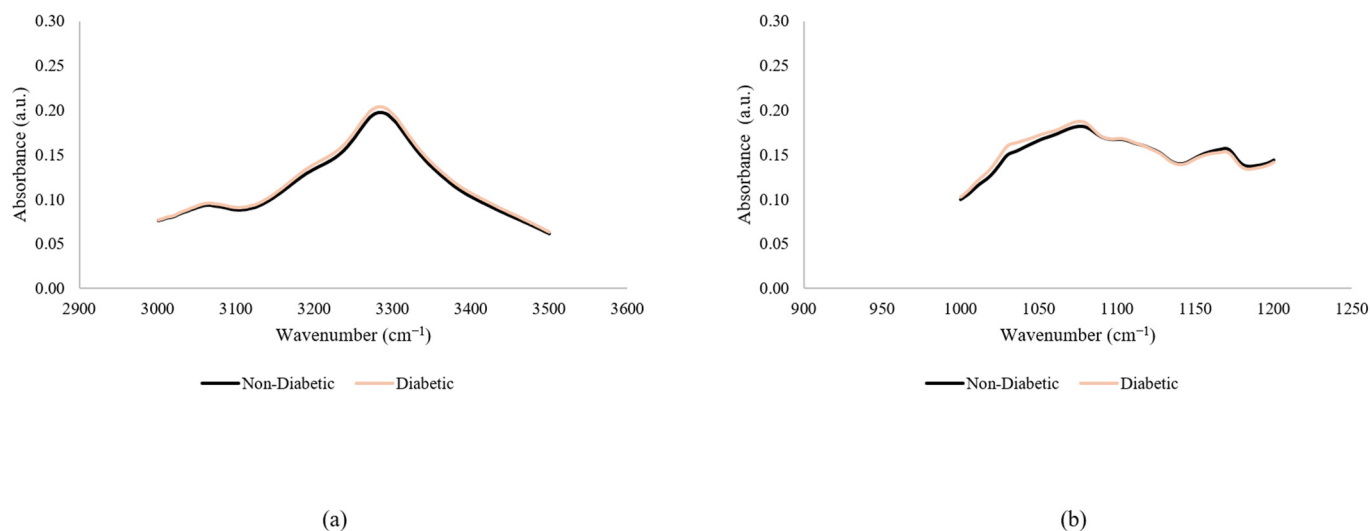


Figure 10. FTIR absorbance spectra for the wavenumber intervals: (a) 3000–3500 cm^{-1} , (b) 1000–1200 cm^{-1} , which represent the peak magnitude differences between the non-diabetic and diabetic spectra.

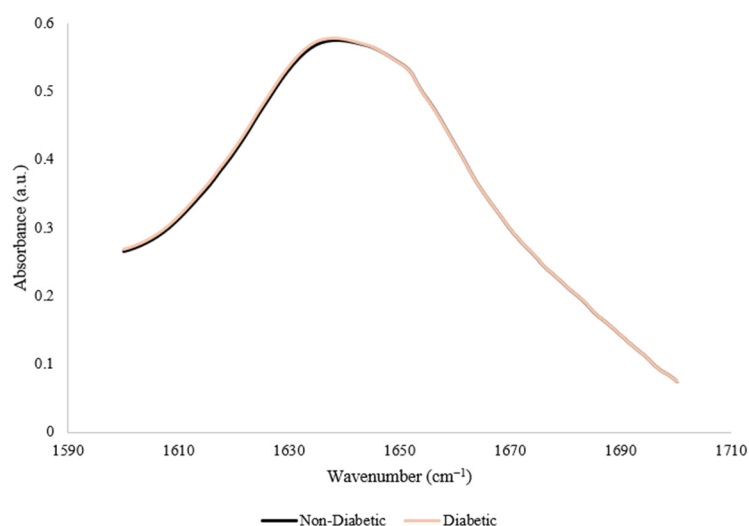


Figure 11. FTIR absorbance spectra for the region 1650 cm^{-1} , which represents a slight peak shift difference between the non-diabetic and diabetic spectra.

In FTIR spectroscopy, spectral differentiation is not based on a single criterion; rather, multiple analytical approaches can be employed depending on whether the comparison is visual, chemical, or statistical in nature [76]. Common methods include peak position analysis, peak intensity comparison, peak area integration, band ratio evaluation, difference spectroscopy, derivative spectroscopy, and multivariate statistical analysis [76]. In the present study, spectral interpretation primarily focused on variations in peak position (i.e., wavenumber shifts) and peak absorbance (intensity). Shifts in characteristic absorption bands may indicate underlying chemical or structural modifications within the samples. As shown in Figure 11, slight positional shifts were observed in the $\sim 1650 \text{ cm}^{-1}$ region, corresponding to the Amide I band, which is associated with C=O stretching vibrations of proteins [77]. This region is widely recognized as indicative of protein secondary structure (see Table 6). The non-diabetic spectrum (black) exhibited slightly lower absorbance value than the diabetic spectrum (orange), suggesting potential alterations in protein conformation. In Figure 10a,b, the 1000–1200 cm^{-1} region, commonly referred to as the

carbohydrate region, displayed more pronounced intensity variations [78]. This spectral range corresponds to C–O and C–O–C stretching vibrations associated with glucose and polysaccharides (Table 6). Given that DM is characterized by altered glucose metabolism, variations in this region are particularly significant. In the 3000–3500 cm^{-1} region, non-diabetic samples demonstrated slightly lower absorbance values and minor positional variations compared to diabetic samples. This region is associated with CH_2 and CH_3 stretching vibrations linked to lipid content [79]. The diabetic spectra exhibited a subtle increase in absorbance peak and/or shifts (Figure 10a), suggesting diabetes related biochemical alterations. These observations are consistent with known metabolic and structural changes associated with glucose dysregulation. These trends were observed in the spectra generated by both models, GAN and AE.

Table 6. General FTIR spectral regions commonly observed in biological samples (specifically blood).

Wavenumber (cm^{-1})	Functional Group	Biomolecule
1000–1100	C–O/C–O–C	Carbohydrates
1240–1300	Amide III	Proteins
1450	CH_2 bending	Lipids
1540	Amide II	Proteins
1650	Amide I (C=O)	Proteins
1740	C=O (ester)	Lipids
2850–2950	C–H stretch	Lipids
3200–3400	O–H/N–H	Water/Proteins

Overall, both spectral and correlation analyses indicate that the GAN and AE models effectively learned and replicated the underlying data distribution. The synthetic spectra closely matched the original infrared signatures of diabetic and non-diabetic samples, particularly in the protein ($\sim 1650 \text{ cm}^{-1}$) and carbohydrate (1000–1200 cm^{-1}) regions. These findings demonstrate that both generative models can capture disease-specific biochemical variations while preserving the fundamental spectral characteristics of the original dataset. Although the AE-generated and GAN-generated spectra appear visually similar in Figures 8 and 9, visual inspection alone does not fully capture the high-dimensional statistical structure of the FTIR dataset. Each spectrum originally contains approximately 3000 wavenumber variables, and subtle differences distributed across many spectral regions may remain imperceptible in graphical representations while still being detectable by ML classifiers. The AE model was trained to minimize reconstruction errors and therefore tends to preserve subtle sample-specific characteristics present in the original dataset. Consequently, classifiers achieved relatively high accuracy and AUC values in distinguishing AE-generated spectra from original spectra. In contrast, the GAN model was optimized via adversarial learning to reproduce the overall data distribution and generate samples that are difficult to distinguish from the original observations. As a result, classifier performance approached random guessing, leading to lower accuracy, ROC, and AUC values. Therefore, the observed differences in classifier performance reflect differences in the underlying statistical distributions learned by the generative models rather than obvious visual differences in the spectral profiles themselves.

6.1.7. Correlation Analysis

This subsection presents a comparative analysis of synthetic and original non-diabetic whole blood FTIR spectra obtained from experimental measurements to evaluate the ability of the generative models to reproduce realistic spectral data. Using correlation metrics, the study assesses the extent to which spectra generated by the GAN and AE models

capture the statistical and structural characteristics of the original samples. By examining the relationship between synthetic and original datasets, this analysis aims to validate the effectiveness of both models in learning and reproducing the underlying spectral distribution. Figure 12 presents correlation plots comparing synthetic non-diabetic FTIR spectra generated by the GAN and AE models with the corresponding experimentally acquired original non-diabetic whole blood spectra.

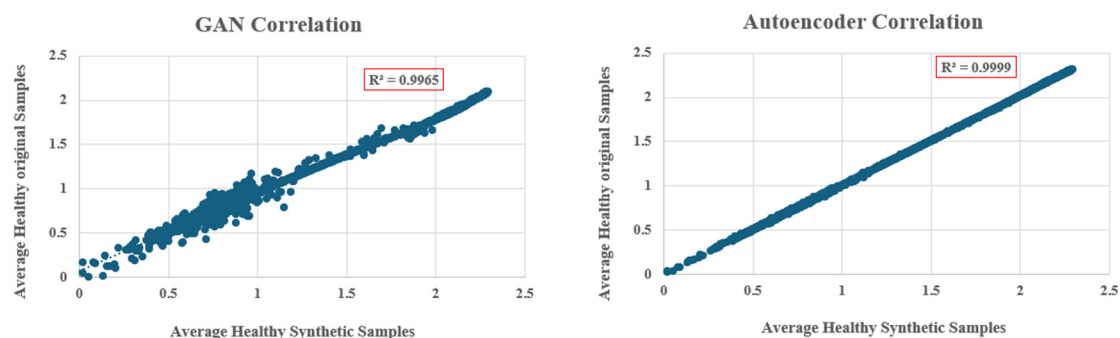


Figure 12. Correlation plots comparing synthetic non-diabetic FTIR spectra with original non-diabetic spectra (experimentally acquired), generated by the GAN and the AE models.

The two scatter plots shown in Figure 13 evaluate the degree of agreement between synthetic average healthy FTIR spectra (x -axis) and the corresponding original average healthy spectra (y -axis). Each point represents the average absorbance intensity at a specific wavenumber across the analyzed spectra. In both cases, a strong linear relationship is observed between the synthetic and original spectral values. The pronounced diagonal trend indicates that the generated spectra quantitatively approximate the original spectra across the full absorbance range, suggesting that both generative models successfully captured the principal characteristics of the underlying data distribution. To quantitatively assess the similarity between original and synthetic spectra, a correlation analysis was performed using the outputs of the GAN and AE models. The dataset comprised six samples with 3601 wavelength variables, and wavelength average spectra were calculated prior to analysis. The resulting correlation metrics are summarized in Table 7.

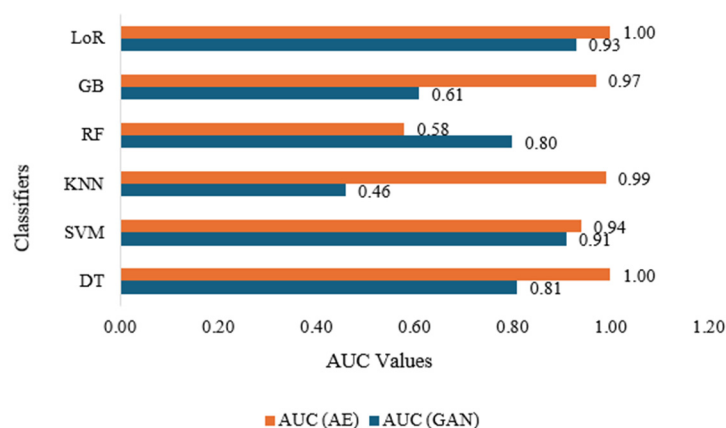


Figure 13. Comparisons between the AUC values of the non-diabetic and the diabetic for the GAN and the AE models.

Table 7. Correlation (r) and Coefficient of Determination (R^2) calculated for the GAN and AE models.

Metric	Comparison	GAN	AE
--------	------------	-----	----

Correlation Coefficient (r)	Original Samples vs Synthetic Samples	0.9982	0.9990
Coefficient of Determination (R^2)	Original Samples vs Synthetic Samples	0.9965	0.9999

The correlation analysis demonstrated excellent agreement between the original and synthetic spectra for both generative models. Specifically, the GAN model achieved $r = 0.9982$ and $R^2 = 0.9965$, whereas the AE model achieved $r = 0.9990$ and $R^2 = 0.9999$. The mean R^2 value across both models was 0.9982, calculated from the GAN R^2 value of 0.9965 and the AE R^2 value of 0.9999. This value represents the average coefficient of determination across the two models and should not be interpreted as the individual R^2 value of either model. These results indicate that both generative approaches preserved the principal statistical and spectral characteristics of the original FTIR data with very high fidelity. From a synthetic data realism perspective, spectra that are more difficult to distinguish from the original data are generally considered more realistic. Consequently, the lower classifier accuracy and AUC values observed for GAN-generated spectra should not be interpreted as inferior performance. Instead, they suggest stronger distributional similarity between the generated and original FTIR spectra. In contrast, the higher classifier separability observed for AE-generated spectra indicates that the reconstructed spectra retained characteristics that were easier for classifiers to distinguish from the original data. Therefore, the differences between the GAN and AE models should be interpreted as reflecting distinct generative behaviors rather than indicating that one approach is universally superior to the other.

6.2. Diabetes Classification Performance

This subsection evaluates disease discrimination performance. In this context, higher classification accuracy, sensitivity, specificity, and AUC values indicate better preservation of diabetes related information.

6.2.1. Performance Metrics for Non-Diabetic Versus Diabetic Blood Spectra

The classification results obtained using the synthetic datasets generated by the GAN and AE models demonstrate substantial differences in their ability to preserve class discriminative information between non-diabetic and diabetic blood spectra, as shown in Tables 8–10. For the GAN-generated dataset, the best testing performance was achieved by LoR (85.0% accuracy, 84.75% F1-score), followed closely by SVM, KNN, and DT, which all achieved testing accuracies above 82%. RF achieved moderate accuracy (75%), whereas GB achieved the lowest (61.67%). The corresponding 10-fold cross-validation results showed relatively consistent performance, with SVM, KNN, and RF achieving mean accuracies of 82–85% and moderate standard deviations. These findings suggest that the GAN-generated spectra retain sufficient disease-related spectral information while maintaining realistic sample variability. The similarity between testing and cross-validation results further indicates that the classifiers generalized reasonably well and were not excessively overfitted to the training data. In contrast, the AE-generated dataset produced exceptionally high classification performance. DT and LoR achieved perfect testing accuracy (100%), precision (100%), recall (100%), and F1-score (100%), while KNN and GB achieved testing accuracies of 99% and 97%, respectively. Similarly, the corresponding 10-fold cross-validation results approached or reached perfect classification performance for several classifiers. Although such results may initially appear highly favorable, they require careful interpretation. In practical biomedical applications, perfect classification is rarely observed due to biological variability, measurement noise, and patient heterogeneity. Therefore, the near-perfect performance obtained using AE-generated data likely indicates that the synthetic samples possess highly separable class boundaries and reduced intra-class variability. This behavior suggests that the AE model may have generated spec-

tra that are easier to classify than genuine biological data, thereby reducing the classification task's complexity. Consequently, the observed improvement in classification accuracy may not correspond to greater realism or a better representation of the underlying biological distributions. Instead, it may reflect an oversimplification of the data structure introduced during synthetic data generation. Overall, the diabetic classification results indicate that both generative approaches successfully preserved disease-related information. However, the GAN-generated dataset appears to provide a more realistic representation of class variability, whereas the AE-generated dataset produces overly separable classes that facilitate classification but may not accurately reflect real-world biological complexity.

Table 8. Test-set classification performance metrics for distinguishing non-Diabetic and Diabetic FTIR spectra in the mixed dataset.

GAN					AE				
Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
LoR	85.00	86.21	83.33	84.75	DT	100.00	100.00	100.00	100.00
SVM	82.50	88.24	75.00	81.08	LoR	100.00	100.00	100.00	100.00
KNN	82.50	88.24	75.00	81.08	KNN	99.00	98.04	100.00	99.01
DT	82.50	84.21	80.00	82.05	GB	97.00	100.00	94.00	96.91
RF	75.00	77.78	70.00	73.68	SVM	91.25	90.24	92.50	91.36
GB	61.67	62.96	56.67	59.65	RF	51.67	51.11	76.67	61.33

Table 9. Training-set classification performance metrics for distinguishing non-diabetic and diabetic FTIR spectra in the mixed dataset.

GAN					AE				
Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
RF	98.12	1.00	96.25	98.09	DT	100.00	100.00	100.00	100.00
KNN	93.75	92.68	95.00	93.83	LoR	100.00	100.00	100.00	100.00
SVM	91.87	93.51	90.00	91.72	GB	98.00	96.15	100.00	98.04
DT	90.00	93.24	86.25	89.61	KNN	96.00	92.59	100.00	96.15
LoR	80.71	82.09	78.57	80.29	SVM	89.17	85.07	95.00	89.76
GB	72.14	79.25	60.00	68.29	RF	72.86	67.78	87.14	76.25

Table 10. Stratified 10-fold cross-validation performance metrics for non-diabetic and diabetic FTIR spectra discrimination.

GAN					AE				
Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	85.00 ± 0.06	87.63 ± 0.09	83.00 ± 0.11	84.44 ± 0.07	LoR	100.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	100.00 ± 0.00
KNN	84.50 ± 0.06	85.96 ± 0.09	84.00 ± 0.08	84.47 ± 0.06	DT	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
RF	82.50 ± 0.09	85.09 ± 0.11	80.00 ± 0.12	81.92 ± 0.09	KNN	98.00 ± 0.01	96.28 ± 0.03	100.00 ± 0.00	98.07 ± 0.01
LoR	79.50 ± 0.06	82.17 ± 0.10	78.00 ± 0.14	78.64 ± 0.07	GB	98.00 ± 0.01	98.05 ± 0.02	98.00 ± 0.02	98.00 ± 0.01
DT	71.00 ± 0.09	74.34 ± 0.13	67.00 ± 0.11	69.73 ± 0.09	SMV	90.50 ± 0.08	89.82 ± 0.08	92.00 ± 0.09	90.61 ± 0.08
GB	66.00 ± 0.04	71.06 ± 0.07	58.00 ± 0.15	62.30 ± 0.07	RF	57.00 ± 0.12	56.54 ± 0.12	64.00 ± 0.20	58.91 ± 0.13

6.2.2. Area Under the Curve of Non-Diabetic Versus Diabetic Blood Spectra

The AUC provides a threshold-independent measure of classifier performance and quantifies a model's ability to discriminate between diabetic and non-diabetic samples. An AUC value of 1.0 indicates perfect discrimination, whereas an AUC value of 0.5 corresponds to random classification with no discriminative capability. Figure 13 presents

the AUC values obtained for the six evaluated ML classifiers using datasets augmented with synthetic spectra generated by the GAN and AE models. Overall, both synthetic data generation approaches produced classifiers with strong discriminative capability; however, the AE-generated dataset consistently yielded superior performance across all evaluated classifiers. For the AE-generated dataset, AUC values ranged from 0.94 to 1.00, indicating excellent to near-perfect classification performance. DT and LoR achieved the highest AUC values of 1.00, followed by KNN (0.99), GB (0.97), and SVM (0.94). These results demonstrate that the AE-generated spectra preserve disease-related spectral characteristics with high fidelity, enabling accurate separation of diabetic and non-diabetic samples across multiple classification algorithms. The GAN-generated dataset also achieved strong classification performance, though AUC values were generally lower and more variable across classifiers. AUC values ranged from 0.46 to 0.93, with LoR achieving the highest value (0.93), followed by SVM (0.91), DT (0.81), RF (0.80), GB (0.61), and KNN (0.46). While several classifiers maintained good discriminatory capability, the substantial reduction in AUC observed for GB, and particularly KNN, suggests that some disease-related discriminative information was less consistently preserved in the GAN-generated dataset. A comparison between the two generative approaches reveals that AE-generated data provided a more stable and informative feature representation for diabetes classification. The consistently high AUC values across all classifiers indicate that the AE effectively preserved the dominant spectral patterns of diabetic and non-diabetic samples. This behavior is consistent with the AE's reconstruction learning mechanism, which emphasizes retaining the most informative latent features present in the original spectra. Consequently, the resulting synthetic samples maintain strong disease-related discriminative characteristics that can be readily exploited by downstream classifiers.

In contrast, the GAN model was primarily optimized to reproduce the overall statistical distribution of the original spectra rather than to maximize class separability. As a result, although the GAN-generated samples exhibited strong realism characteristics, some disease-specific discriminative features may have been partially smoothed or redistributed during adversarial training, leading to lower classification performance for certain classifiers. These findings highlight the trade-off between synthetic data realism and disease discrimination capability. While GAN-generated spectra may more closely resemble the overall distribution of original FTIR data, AE-generated spectra appear to preserve disease-related information more effectively, resulting in superior diabetes classification performance. Collectively, the AUC analysis demonstrates that both generative models can support diabetes classification; however, the AE-generated dataset consistently achieved higher and more stable discrimination performance across all classifiers. These results suggest that synthetic augmentation of AE may be particularly advantageous when the primary objective is to maximize classification accuracy while preserving disease-related spectral information.

6.2.3. Comparative Interpretation of GAN and AE Models

The AE model consistently produced higher diabetic classification accuracies than the GAN model, suggesting that the generated spectra contained highly separable disease-related features. However, the realism assessment revealed that AE-generated samples could be distinguished from original samples with relatively high accuracy, indicating reduced fidelity to the original spectral distribution. Conversely, the GAN model produced more moderate diabetic classification performance but substantially lower original versus synthetic accuracy. This finding indicates that GAN-generated samples more closely resembled the original FTIR spectra and preserved realistic spectral variability. Therefore, from a synthetic data quality perspective, the GAN model appears to generate more realistic spectra, whereas the AE model appears to generate spectra that facilitate

classification but may not accurately reflect the complexity of original biological samples. The combined results demonstrate that both generative approaches successfully generated synthetic FTIR blood spectra capable of supporting diabetes classification. However, the two models exhibited fundamentally different characteristics. The GAN model generated highly realistic spectra that were difficult to distinguish from original samples, indicating superior fidelity to the underlying data distribution. In contrast, the AE model generated spectra that produced near-perfect diabetic classification performance but remained relatively easy to distinguish from original spectra, suggesting reduced realism and potentially inflated classification performance. Consequently, while AE-generated data may improve classification accuracy, GAN-generated data appear to provide a more reliable representation of real-world spectral variability and therefore constitute a more suitable choice for privacy-preserving synthetic data generation in FTIR diabetes research.

6.3. Preliminary Experimental Validation and Coverage Analysis

This subsection presents the experimental validation of the proposed framework using independently acquired whole blood FTIR spectra. Unlike the previous analyses, which focused on synthetic data realism and diabetes discrimination using the original and generated datasets, this experiment was designed to assess the practical applicability of the framework under a more realistic testing scenario. To assess the translational feasibility of the proposed serum synthetic data framework, a preliminary experimental validation was conducted using independently acquired whole blood FTIR spectra. This component was designed as an exploratory feasibility assessment rather than a definitive external clinical validation study. In addition to classification outcomes, a coverage analysis was performed using a 90% posterior probability threshold to quantify the proportion of samples for which the classifiers produced high confidence predictions. The 90% posterior probability threshold was selected as an exploratory confidence criterion to identify predictions associated with relatively high classifier certainty. This threshold was not derived from any regulatory guideline, clinical decision standard, or established diagnostic cutoff. Rather, it was decided to provide a simple, interpretable assessment of prediction confidence in this proof-of-concept feasibility study. Consequently, the reported coverage values should be interpreted as exploratory indicators of classifier confidence rather than validated clinical performance measures. In this context, higher coverage indicates that a larger proportion of samples received confident predictions; however, coverage should not be interpreted as a measure of classification accuracy and should be considered alongside the corresponding prediction accuracy and classification performance metrics.

Specifically, the objective was to assess the effectiveness of the GAN and AE models in generating synthetic data that closely replicate the statistical characteristics of the original spectra, as evaluated through correlation analysis. In addition, the study assessed the performance of multiple ML classifiers in distinguishing between non-diabetic and diabetic samples under a regulatory style confidence constraint. Figure 14 presents the FTIR spectra of human whole blood samples collected during the experiment, which were used to evaluate classifier performance and assess the ability of the generative models to produce synthetic spectra that resemble the original dataset. It should be noted that the experimental validation dataset consisted of independently acquired whole blood samples, whereas the original dataset and the corresponding synthetic data were derived from blood serum spectra. Consequently, differences in absolute absorbance intensity and baseline characteristics are expected because whole blood and serum are biologically distinct sample matrices. Whole blood contains additional cellular constituents and hemoglobin, which are absent from serum and can influence spectral profiles. Furthermore, sample preparation procedures, optical path length, instrumental conditions, and measurement protocols may contribute to differences in absolute spectral magnitudes, baseline inten-

sity, absorbance amplitude, and overall spectral scaling. Accordingly, the purpose of the experimental validation was not to establish direct spectral equivalence, validate absolute spectral amplitudes, or demonstrate one-to-one correspondence with the serum dataset used for model training. Instead, the validation focused on determining whether the principal diabetes related to biochemical absorption regions, characteristic spectral trends, and classifier behavior remained consistent under a more application-oriented sampling condition despite these matrix-related differences. From this perspective, the preservation of characteristic biochemical absorption regions, together with the strong correlation observed between original and synthetic spectra, supports the methodological transferability and preliminary translational relevance of the proposed framework. Consequently, the whole-blood experiment should be interpreted as a preliminary feasibility assessment of spectral transferability and methodological applicability rather than a direct quantitative validation of the serum-generative models or a definitive external clinical validation study.

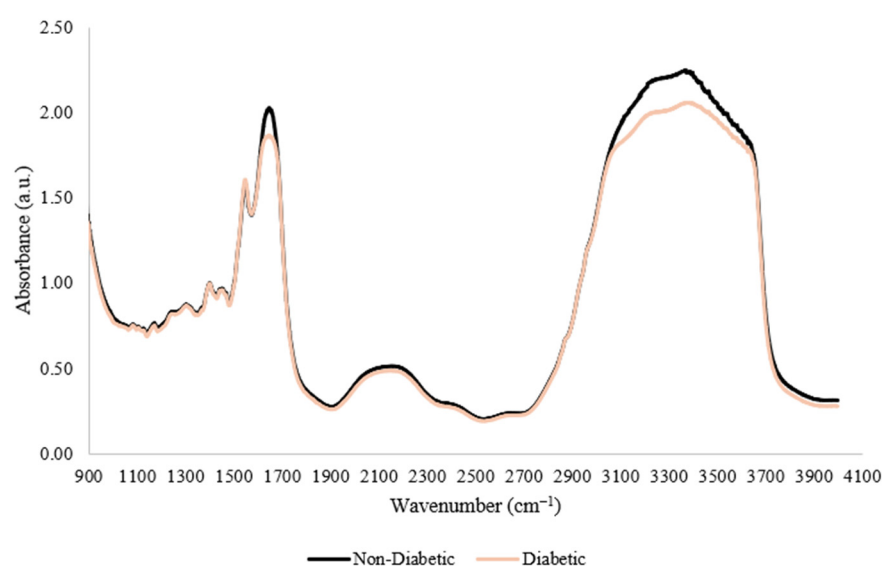


Figure 14. Experimentally acquired Absorbance spectra of the non-diabetic and diabetic whole blood samples overlaid. It compares how infrared light is absorbed by biological samples from non-diabetic vs. diabetic groups across different infrared frequencies.

Performance Assessments of Multiple ML Classifiers

Multiple classifiers were evaluated under an exploratory high-confidence decision framework. Only predictions with a posterior probability $\geq 90\%$ were accepted as automated classifications, whereas predictions below this threshold were assigned to human review. This conservative design was adopted to investigate classifier confidence behavior and automation coverage rather than to validate clinical diagnosis. In a real clinical workflow, such a deferral mechanism could reduce the risk of acting on uncertain AI outputs; however, the present dataset is too small and clinically imbalanced to establish operational performance. Consequently, coverage is reported as the proportion of samples exceeding the exploratory confidence threshold and should be interpreted independently of diagnostic accuracy. Table 11 was evaluated using eight samples: seven experimentally acquired FTIR spectra (six healthy volunteers and one diabetic volunteer) and one additional synthetic diabetic spectrum generated by the GAN model. The synthetic sample was included to assess classifier behavior under a more challenging validation scenario. It is important to note that the preliminary experimental validation was based on FTIR spectra acquired from whole blood samples (without dilution or processing), whereas the

original dataset used for model development consisted of blood serum FTIR spectra. Consequently, whole-blood measurements contained additional biochemical constituents and cellular components, which introduced greater spectral complexity and variability. This increased complexity provided a more demanding assessment of the robustness and confidence behavior of the ML classifiers.

The coverage metric quantifies the proportion of samples for which a classifier produces predictions that satisfy the predefined confidence criterion. Therefore, coverage reflects the applicability of the decision framework rather than the correctness of the classification. A coverage of 100% indicates that all evaluated samples received predictions; however, this metric should be interpreted alongside classification accuracy and related performance measures to assess a model's true predictive capability. Nevertheless, the results demonstrate the potential feasibility of using whole-blood FTIR spectra to distinguish between healthy and pathological samples. This capability is particularly relevant for point-of-care applications, where minimal sample preparation is desirable, and blood filtration, dilution, or component separation may not be required.

Table 11. Classification performance for the 6 classifiers trained on GAN-generated data derived from the original dataset, when applied to experimentally acquired whole blood FTIR spectra.

Classifier	Non-Diabetic	Diabetic	Flag for Human Review	Coverage (%)
LoR	6	2	0	100.0
GB	5	2	1	87.5
KNN	4	1	3	62.5
SVM	2	0	6	25.0
DT	0	0	8	0.0
RF	0	0	8	0.0

The results showed that the DT and RF classifiers rejected all samples, resulting in 0% coverage. This finding indicates that none of their predicted probabilities exceeded the predefined confidence threshold. Such behavior suggests either conservative probability estimation or insufficient class separation within the current dataset. Although this outcome maximizes safety by preventing potentially uncertain automated decisions, it substantially limits automation capability. The SVM classifier exhibited very low coverage (25%) and confidently classified only 2 samples. This behavior is consistent with the conservative confidence estimates often produced by SVM models, particularly when trained on small datasets. While such behavior may enhance safety, it also restricts the proportion of samples that can be automatically classified. KNN achieved moderate coverage (62.5%), indicating that it was confident for a subset of samples. Because KNN relies on local neighborhood relationships, its confidence can vary considerably depending on the distribution and complexity of the available data. GB demonstrated high coverage (87.5%), indicating strong nonlinear class separation and well-defined probability outputs. This suggests that the model effectively captured complex patterns in the data while maintaining high-confidence predictions. LoR achieved full coverage (100%), classifying all samples above the predefined confidence threshold. This result indicates strong separability in the learned feature space and consistently high posterior probabilities. Although high coverage improves automation capability, further calibration analysis is required to ensure that probability estimates are not overly confident. The classifiers are ordered by coverage percentage, which reflects the proportion of samples for which predictions exceed the predefined confidence threshold. This ranking emphasizes practical applicability, as higher coverage indicates the classifier's greater ability to provide confident predictions under real experimental conditions. These findings show that classifier performance under regulatory-style confidence constraints depends not only on discriminative accuracy but also on

probability calibration and confidence behavior. Among the evaluated models, GB and LoR demonstrated the most favorable balance between safety and automation, whereas the tree-based models (DT and RF) were overly conservative under the strict confidence threshold. These results highlight the importance of evaluating coverage and confidence dynamics when designing clinically deployable AI systems. Overall, the developed classifiers successfully distinguished between non-diabetic and diabetic samples under a high-confidence decision framework. Preliminary findings demonstrate discriminative capability across multiple classifiers; however, due to the limited sample size, further validation in larger, independent cohorts is required before clinical applicability can be established. Overall, both generative models successfully captured the intrinsic structure of the data and demonstrated strong capability in generating realistic synthetic spectra. Correlation analysis confirmed that the GAN and AE models effectively learned the underlying spectral distribution and generated highly realistic synthetic spectra, achieving a mean R^2 of 0.9982, indicating near-perfect agreement with the original data. In parallel, the implementation of a regulatory-style high-confidence decision framework enabled the evaluation of classifier reliability under strict clinical constraints. The results revealed notable differences in classifier automation capacity, with GB and LoR providing the most favorable balance between diagnostic confidence and automated decision coverage. Taken together, these findings highlight the potential of combining synthetic data augmentation with confidence-aware ML models to support reliable FTIR diabetes detection. Nevertheless, the objective of the coverage analysis was to evaluate classifier confidence behavior rather than diagnostic performance. Because the preliminary experimental dataset consisted of only seven independently acquired whole-blood spectra and one additional synthetic diabetic spectrum, the sample size was insufficient to support statistically meaningful estimates of classification accuracy, sensitivity, specificity, or false-positive rates. Consequently, coverage was reported as an exploratory indicator of automation capability, whereas future studies involving larger independent cohorts will be required to assess predictive accuracy, probability calibration, robustness, and clinical applicability under representative real-world conditions.

6.4. Implications of the Main Findings and Preliminary Whole Blood Validation

Compared with prior FTIR-based studies for diabetes analysis, which have mainly focused on direct classification of measured spectra, the present work contributes a more explicitly data-centric and privacy-aware framework based on synthetic data generation from FTIR blood serum spectra. The main contribution of the study is the comparative evaluation of AE and GAN in terms of spectral fidelity, downstream classification utility, and realism-oriented augmentation. The results show a clear trade-off between the two approaches: AE-generated data preserved a stronger class discriminative structure and yielded better downstream classification performance, whereas GAN-generated data were generally harder to distinguish from the original data, suggesting greater realism and potential value for privacy-preserving augmentation. The spectral analyses further support these findings by showing that both models retained diagnostically relevant biochemical variations, particularly in the Amide I, carbohydrate, and lipid-associated regions. Within this context, the whole blood experiment should be interpreted as a preliminary validation step rather than as a second full-scale study or a definitive external validation cohort. Its purpose is to provide initial evidence that the proposed framework may remain relevant under a more practical and less processed sampling condition, despite the added biochemical complexity of whole blood relative to serum. Therefore, the study's central innovation remains the serum-based synthetic data framework and its comparative methodological assessment, while the whole blood results strengthen the translational relevance of the study in an exploratory manner. Future work should extend these

findings through larger independent cohorts, external validation, privacy risk assessment, and calibration-aware evaluation to confirm robustness and clinical applicability.

7. Conclusions and Future Work

This study investigated the feasibility of privacy-preserving synthetic data generation for FTIR diabetes classification using two deep generative approaches: AE and GAN. Using a dataset of 100 blood serum FTIR spectra, both models successfully generated synthetic spectra that preserved the principal statistical, structural, and biochemical characteristics of the original data. Correlation analysis demonstrated very high agreement between the original and generated spectra, confirming the ability of both generative approaches to reproduce the underlying spectral distribution with high fidelity. The comparative evaluation revealed that AE-generated spectra exhibited higher classifier separability, indicating that the reconstructed spectra retained stronger discriminative characteristics from the original dataset. In contrast, GAN-generated spectra were more difficult for classifiers to distinguish from original spectra, suggesting a closer alignment with the original data distribution and greater realism from a synthetic data generation perspective. These findings highlight that classifier separability and synthetic data quality represent distinct evaluation objectives and should not be interpreted interchangeably. Consequently, higher classification performance in the realism assessment task does not necessarily indicate superior synthetic data quality but rather reflects differences in the generative characteristics of the models. The analysis further demonstrated that the most prominent diabetes spectral variations were concentrated within the carbohydrate-associated ($1000\text{--}1200\text{ cm}^{-1}$), Amide I ($\sim 1650\text{ cm}^{-1}$), and lipid-associated ($3000\text{--}3500\text{ cm}^{-1}$) regions. These findings support the potential of FTIR spectroscopy combined with synthetic data augmentation and ML as a promising framework for privacy-aware biomedical data analysis and diabetes screening applications. A preliminary experimental feasibility assessment was also conducted using independently acquired whole blood FTIR spectra. Although differences in spectral intensity were observed between serum and whole blood measurements, the principal biochemical absorption regions remained consistent. Because the experimental cohort included only one diabetic volunteer, this analysis should be interpreted solely as a proof-of-concept assessment of spectral feasibility and methodological transferability rather than validation of diabetes classification performance or clinical diagnostic capability.

A limitation of the present study is the relatively modest size of the original FTIR blood serum dataset (100 spectra) used to train the generative models. Although this sample size is comparable to those reported in several FTIR diabetes studies, it remains limited for training deep generative architectures and may not fully capture the biological and spectral variability present in larger and more diverse populations. Consequently, the generated synthetic spectra should be interpreted as approximations of the underlying data distribution rather than comprehensive representations of all potential clinical variability. In addition, a formal statistical power analysis was not performed because the primary objective of this work was the methodological evaluation of synthetic data generation, synthetic data realism, and classification feasibility rather than hypothesis testing. Therefore, the findings should be regarded as a proof-of-concept demonstration of synthetic FTIR data augmentation and privacy-preserving ML analysis rather than definitive evidence of clinical effectiveness or broadly generalizable diagnostic performance. Future studies should evaluate the proposed framework using larger, independent, and multi-center cohorts comprising both serum and whole blood samples to establish robustness, external validity, and clinical applicability. Future work will focus on validating the proposed framework using larger and more diverse FTIR datasets obtained from multiple institutions and independent patient cohorts. Additional studies should investigate ad-

vanced synthetic data evaluation metrics, diversity preservation measures, and alternative generative architectures capable of further improving realism and robustness. Future research should also explore the generation of synthetic pathological spectra using larger diabetic cohorts, assess model performance under external validation protocols, and evaluate robustness across different FTIR acquisition systems and biological sample types. Finally, prospective studies integrating synthetic data generation, FTIR spectroscopy, and ML decision support into clinically relevant workflows will be essential to assess regulatory, operational, and translational requirements for real-world healthcare deployment.

Author Contributions: Conceptualization: A.F.; S.O.C.; G.M.; and V.C.; methodology: A.F.; H.T.; S.O.C.; G.M.; and V.C.; software: A.F.; H.T.; S.O.C.; G.M.; and V.C.; validation: A.F.; K.N.B.; M.G.; H.T.; S.O.C.; G.M.; and V.C.; formal analysis: A.F.; investigation: A.F.; H.T.; S.O.C.; G.M.; and V.C.; resources: K.N.B.; M.G.; S.O.C.; G.M.; and V.C.; data curation: A.F.; K.N.B.; M.G.; H.T.; S.O.C.; G.M.; and V.C.; writing—original draft preparation: A.F.; H.T.; S.O.C.; G.M.; and V.C.; writing—review and editing: A.F.; K.N.B.; M.G.; H.T.; S.O.C.; G.M.; and V.C.; visualization: A.F.; H.T.; S.O.C.; G.M.; and V.C.; supervision: S.O.C.; G.M.; and V.C.; project administration: S.O.C.; G.M.; and V.C.; funding acquisition: S.O.C.; G.M.; and V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the CardioSensing (COMPETE2030-FEDER-00689800), reference 2023.16645.ICDT (DOI: 10.54499/2023.16645.ICDT), and by the NeuroTech4Med, reference COMPETE2030-FEDER-00712000 projects, co-financed by European Regional Development Fund (FEDER) and national funds, through Fundação para a Ciência e a Tecnologia (FCT, I.P.), through the Innovation and Digital Transition Programme—Compete2030. We also acknowledge the partial financial support by national funds of FCT, under the support to R&D units grant, through the reference projects doi.org/10.54499/UID/04436/2025 (CMEMS), LA/P/0029/2020 (LABBELS), UIDB/05549:2Ai and UIDP/05549:2Ai. A.F. thanks FCT for his 2023.03312.BD PhD grant.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee for Research in Life and Health Sciences (CEICVS), University of Minho, which issued a favorable opinion for the project “Fabrico de um novo dispositivo de diagnóstico in vitro baseado em espectroscopia FTIR para análise de sangue humano” (Approval Code: CEICVS 039/2023; Approval Date: 12 July 2023).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. Whole blood samples were collected from adult volunteers, and only gender, age, and disease status (healthy/pathological) were recorded. No additional participant-identifying information was collected.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors thank Maria Conceição Paiva and the Institute of Polymers and Composites (IPC, University of Minho) for providing access to FTIR equipment and facilities.

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

References

1. Croft, P.; Altman, D.G.; Deeks, J.J.; Dunn, K.M.; Hay, A.D.; Hemingway, H.; LeResche, L.; Peat, G.; Perel, P.; Petersen, S.E.; et al. The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about “what is likely to happen” should shape clinical practice. *BMC Med.* **2015**, *13*, 20. <https://doi.org/10.1186/s12916-014-0265-4>. PMID: 25637245; PMCID: PMC4311412.
2. Phillips, C.; Parkinson, A.; Namsrai, T.; Chalmers, A.; Dews, C.; Gregory, D.; Kelly, E.; Lowe, C.; Desborough, J. Time to diagnosis for a rare disease: Managing medical uncertainty. A qualitative study. *Orphanet J. Rare Dis.* **2024**, *19*, 297. <https://doi.org/10.1186/s13023-024-03319-2>. PMID: 39143641; PMCID: PMC11323401.

3. Moroz, T.; Taylor, M.; Schmitt, P.; Donovan, D.J.; Nori, M. How many zebras are there now? An updated report on publications on rare diseases in medical literature. *Future Rare Dis.* **2025**, *5*, 2499402. <https://doi.org/10.1080/23995270.2025.2499402>.
4. Chandan, R.R.; Singh, J.; Ravi, V.; Shivahare, B.D.; Alahmadi, T.J.; Singh, P.; Diwakar, M. Reviewing the Impact of Machine Learning on Disease Diagnosis and Prognosis: A Comprehensive Analysis. *Open Pain J.* **2024**, *17*, e18749445297804, <https://doi.org/10.2174/0118763863291395240516093102>.
5. Ahsan, M.; Mahmud, M.A.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies* **2021**, *9*, 52. <https://doi.org/10.3390/technologies9030052>.
6. Lyu, G. Data-driven decision making in patient management: A systematic review. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 239. <https://doi.org/10.1186/s12911-025-03072-x>.
7. Qasrawi, R.; Issa, G.; Thwib, S.; AbuGhoush, R.; Amro, M.; Ayyad, R.; Vicuna, S.; Badran, E.; Khader, Y.; Al Qutob, R.; et al. The role of machine learning in infectious disease early detection and prediction in the MENA region: A systematic review. *Inform. Med. Unlocked* **2025**, *56*, 101651. <https://doi.org/10.1016/j.imu.2025.101651>.
8. Olabiyi, W.; Akinyele, D. The Evolution of AI: From Rule-Based Systems to Data-Driven Intelligence. 2025. Available online: <https://www.researchgate.net/publication/388035967> (accessed on 15 January 2025).
9. Fazil, A.W. Artificial Intelligence in Infectious Disease Detection and Control: A Systematic Literature Review of Advancements, Challenges, and Future Directions. *Afghan. J. Infect. Dis.* **2025**, *3*, 113–126. <https://doi.org/10.60141/ajid.96>.
10. Murtaza, H.; Ahmed, M.; Khan, N.F.; Murtaza, G.; Zafar, S.; Bano, A. Synthetic data generation: State of the art in health care domain. *Comput. Sci. Rev.* **2023**, *48*, 100546. <https://doi.org/10.1016/j.cosrev.2023.100546>.
11. Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Science* **2024**, *6*, 3. <https://doi.org/10.3390/sci6010003>.
12. Abouelmehdi, K.; Beni-Hessane, A.; Khaloufi, H. Big healthcare data: Preserving security and privacy. *J. Big Data* **2018**, *5*, 1. <https://doi.org/10.1186/s40537-017-0110-7>.
13. Conduah, A.K.; Ofoe, S.; Siaw-Marfo, D. Data privacy in healthcare: Global challenges and solutions. *Digit. Health* **2025**, *11*, 20552076251343959. <https://doi.org/10.1177/20552076251343959>.
14. World Health Organization. The Protection of Personal Data in Health Information Systems-Principles and Processes for Public Health. Available online: <https://iris.who.int/server/api/core/bitstreams/2f9bfa5b-4fcd-4553-a211-6ea19686ea4d/content> (accessed on 19 March 2021).
15. McGuckin, T.; Crick, K.; Myroniuk, T.W.; Setchell, B.; Yeung, R.O.; Campbell-Scherer, D. Understanding challenges of using routinely collected health data to address clinical care gaps: A case study in Alberta, Canada. *BMJ Open Qual.* **2022**, *11*, e001491. <https://doi.org/10.1136/bmjoq-2021-001491>. PMID: 34996811; PMCID: PMC8744094.
16. Alhumaidi, N.H.; Dermawan, D.; Kamaruzaman, H.F.; Alotaqi, N. The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review. *JMIR Med. Inform.* **2025**, *13*, e68898. <https://doi.org/10.2196/68898>.
17. Tsopra, R.; Fernandez, X.; Luchinat, C.; Alberghina, L.; Lehrach, H.; Vanoni, M.; Dreher, F.; Sezerman, O.; Cuggia, M.; de Tayrac, M.; et al. A framework for validating AI in precision medicine: Considerations from the European ITFoC consortium. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 274. <https://doi.org/10.1186/s12911-021-01634-3>.
18. Ahmed, M.M.; Okesanya, O.J.; Oweidat, M.; Othman, Z.K.; Musa, S.S.; Iii, D.E.L.-P. The ethics of data mining in healthcare: Challenges, frameworks, and future directions. *BioData Min.* **2025**, *18*, 47. <https://doi.org/10.1186/s13040-025-00461-w>.
19. Bustos, M.L.; Song, K.; Brochu, H.N.; Zhang, Q.; Iyer, L.K.; Icenhour, C.R. Impact of non-standardized reporting on reproducibility, usability, and integration in nasopharyngeal metagenomic research: A systematic review. *Front. Microbiol.* **2026**, *17*, 1707004. <https://doi.org/10.3389/fmicb.2026.1707004>.
20. El Arab, R.A.; Abu-Mahfouz, M.S.; Abuadas, F.H.; Alzghoul, H.; Almari, M.; Ghannam, A.; Seweid, M.M. Bridging the Gap: From AI Success in Clinical Trials to Real-World Healthcare Implementation—A Narrative Review. *Healthcare* **2025**, *13*, 701. <https://doi.org/10.3390/healthcare13070701>.
21. Gonzales, A.; Guruswamy, G.; Smith, S.R. Synthetic data in health care: A narrative review. *PLoS Digit. Health* **2023**, *2*, e0000082. <https://doi.org/10.1371/journal.pdig.0000082>.
22. Ennab, M.; Mcheick, H. Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. *Front. Robot. AI* **2024**, *11*, 1444763. <https://doi.org/10.3389/frobt.2024.1444763>.
23. Jenkins, D.A.; Martin, G.P.; Sperrin, M.; Riley, R.D.; Debray, T.P.A.; Collins, G.S.; Peek, N. Continual updating and monitoring of clinical prediction models: Time for dynamic prediction systems? *Diagn. Progn. Res.* **2021**, *5*, 1. <https://doi.org/10.1186/s41512-020-00090-3>.

24. Tasnia, R.; Obeng-Gyasi, E. Evaluating Machine Learning Models for Classifying Diabetes Using Demographic, Clinical, Life-style, Anthropometric, and Environmental Exposure Factors. *Toxics* **2026**, *14*, 76. <https://doi.org/10.3390/toxics14010076>.
25. Yang, J.; Soltan, A.A.S.; Clifton, D.A. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *npj Digit. Med.* **2022**, *5*, 69. <https://doi.org/10.1038/s41746-022-00614-9>.
26. Koul, A.; Duran, D.; Hernandez-Boussard, T. Synthetic data, synthetic trust: Navigating data challenges in the digital revolution. *Lancet Digit. Health* **2025**, *7*, 100924. <https://doi.org/10.1016/j.landig.2025.100924>.
27. Fahim, Y.A.; Hasani, I.W.; Kabba, S.; Ragab, W.M. Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. *Eur. J. Med. Res.* **2025**, *30*, 848. <https://doi.org/10.1186/s40001-025-03196-w>.
28. Ayoub, M.; Zhao, H.; Li, L.; Yang, D.; Hussain, S.; Wahid, J.A. Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. *Commun. Med.* **2026**, *6*, 86. <https://doi.org/10.1038/s43856-025-01348-x>.
29. Almehwari, S.A.; Almalki, I.S.; Abumilha, B.A.; Altharwi, B.H. Improving Hospital Efficiency and Cost Management: A Systematic Review and Meta-Analysis. *Cureus* **2024**, *16*, e71721. <https://doi.org/10.7759/cureus.71721>.
30. Noguer, J.; Contreras, I.; Mujahid, O.; Beneyto, A.; Vehi, J. Generation of Individualized Synthetic Data for Augmentation of the Type 1 Diabetes Data Sets Using Deep Learning Models. *Sensors* **2022**, *22*, 4944. <https://doi.org/10.3390/s22134944>.
31. Ghosh, K.; Chandra, S.; Ghosh, S.; Ghosh, U.S. Artificial Intelligence in Personalized Medicine for Diabetes Mellitus: A Narrative Review. *Cureus* **2025**, *17*, e91520. <https://doi.org/10.7759/cureus.91520>.
32. Talukder, A.; Islam, M.; Uddin, A.; Kazi, M.; Khalid, M.; Akhter, A.; Moni, M.A. Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. *Digit. Health* **2024**, *10*, 20552076241271867. <https://doi.org/10.1177/20552076241271867>.
33. Mendes, J.M.; Barbar, A.; Refaie, M. Synthetic data generation: A privacy-preserving approach to accelerate rare disease research. *Front. Digit. Health* **2025**, *7*, 1563991. <https://doi.org/10.3389/fdgth.2025.1563991>.
34. Tabari, P.; Costagliola, G.; De Rosa, M.; Boeker, M. State-of-the-Art Fast Healthcare Interoperability Resources (FHIR)-Based Data Model and Structure Implementations: Systematic Scoping Review. *JMIR Med. Inform.* **2024**, *12*, e58445. <https://doi.org/10.2196/58445>.
35. Yang, B.; Zhang, G.; Wang, K. Transformer-Based Multi-Source Transfer Learning for Intrusion Detection Models with Privacy and Efficiency Balance. *Entropy* **2026**, *28*, 136. <https://doi.org/10.3390/e28020136>.
36. Finetti, R.; Roncaglia, B.; Visibelli, A.; Spiga, O.; Santucci, A. Data Augmentation and Synthetic Data Generation in Rare Disease Research: A Scoping Review. *Med. Sci.* **2025**, *13*, 260. <https://doi.org/10.3390/medsci13040260>.
37. Prendin, F.; Facchinetti, A.; Cappon, G. Data Augmentation Via Digital Twins to Develop Personalized Deep Learning Glucose Prediction Algorithms for Type 1 Diabetes in Poor Data Context. *IEEE Trans. Biomed. Eng.* **2025**, *2025*, 3635264. <https://doi.org/10.1109/tbme.2025.3635264>.
38. Chato, L.; Regentova, E. Survey of Transfer Learning Approaches in the Machine Learning of Digital Health Sensing Data. *J. Pers. Med.* **2023**, *13*, 1703. <https://doi.org/10.3390/jpm13121703>.
39. Lvovs, D.; Creason, A.L.; Levine, S.S.; Noble, M.; Mahurkar, A.; White, O.; Fertig, E.J. Balancing ethical data sharing and open science for reproducible research in biomedical data science. *Cell Rep. Med.* **2025**, *6*, 102080. <https://doi.org/10.1016/j.xcrm.2025.102080>.
40. Li, M.; Tan, A.; Song, L.; Qiu, X.; Zhou, L.; Wen, Z. Performance of diabetes risk prediction models: A systematic review and meta-analysis. *Endocr. Connect.* **2025**, *14*, e250353. <https://doi.org/10.1530/ec-25-0353>.
41. Guang, P.; Huang, W.; Guo, L.; Yang, X.; Huang, F.; Yang, M.; Wen, W.; Li, L. Blood-based FTIR-ATR spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes: A STARD compliant diagnosis research. *Medicine* **2020**, *99*, e19657. <https://doi.org/10.1097/md.00000000000019657>.
42. Yang, T.; Qi, F.; Guo, F.; Shao, M.; Song, Y.; Ren, G.; Linlin, Z.; Qin, G.; Zhao, Y. An update on chronic complications of diabetes mellitus: from molecular mechanisms to therapeutic strategies with a focus on metabolic memory. *Mol Med.* **2024**, *30*, 71. <https://doi.org/10.1186/s10020-024-00824-9>.
43. Colmenares-Mejia, C.C.; García-Suaza, A.F.; Rodríguez-Lesmes, P.; Lochmuller, C.; Atehortúa, S.C.; Camacho-Cogollo, J.; Martínez, J.P.; Rincón, J.; Céspedes, Y.R.; Morales-Mendoza, E.; et al. Predicting diabetes mellitus metabolic goals and chronic complications transitions—Analysis based on natural language processing and machine learning models. *PLoS ONE* **2025**, *20*, e0321258. <https://doi.org/10.1371/journal.pone.0321258>.

44. Endres, M.; Venugopal, A.M.; Tran, T.S. Synthetic Data Generation: A Comparative Study. In Proceedings of the IDEAS'22: International Database Engineered Applications Symposium, Budapest, Hungary, 22–24 August 2022; pp. 94–102. <https://doi.org/10.1145/3548785.3548793>.
45. Miura, T.; Kimura, E.; Ichikawa, A.; Kii, M.; Yamamoto, J. Evaluating Synthetic Data Generation Techniques for Medical Dataset. In Proceedings of the 17th International Conference on Health Informatics, Rome, Italy, 21–23 February 2024; pp. 315–322. <https://doi.org/10.5220/0012314500003657>.
46. Rashidi, H.H.; Albahra, S.; Rubin, B.P.; Hu, B. A novel and fully automated platform for synthetic tabular data generation and validation. *Sci. Rep.* **2024**, *14*, 23312. <https://doi.org/10.1038/s41598-024-73608-0>.
47. Bae, W.D.; Alkobaisi, S.; Horak, M.; Bankar, S.; Bhuvaji, S.; Kim, S.; Park, C.-S. Synthetic Data Generation and Evaluation Techniques for Classifiers in Data Starved Medical Applications. *IEEE Access* **2025**, *13*, 16584–16602. <https://doi.org/10.1109/access.2025.3532222>.
48. Yang, X.; Fang, T.; Li, Y.; Guo, L.; Li, F.; Huang, F.; Li, L. Pre-diabetes diagnosis based on ATR-FTIR spectroscopy combined with CART and XGBoots. *Optik* **2019**, *180*, 189–198. <https://doi.org/10.1016/j.ijleo.2018.11.059>.
49. Dharmarathne, G.; Jayasinghe, T.N.; Bogahawaththa, M.; Meddage, D.; Rathnayake, U. A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Health Anal.* **2024**, *5*, 100301. <https://doi.org/10.1016/j.health.2024.100301>.
50. Barman, D.; Hasnat, A.; Nag, R. An Introduction to Autoencoders. Computation, CSE-GCETTB vol. III. 2022. Available online: https://www.researchgate.net/publication/379542201_AN_INTRODUCTION_TO_AUTOENCODERS (accessed on February 2022).
51. Berahmand, K.; Daneshfar, F.; Salehi, E.S.; Li, Y.; Xu, Y. Autoencoders and their applications in machine learning: A survey. *Artif. Intell. Rev.* **2024**, *57*, 28. <https://doi.org/10.1007/s10462-023-10662-6>.
52. Recanatesi, S.; Farrell, M.; Lajoie, G.; Deneve, S.; Rigotti, M.; Shea-Brown, E. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nat. Commun.* **2021**, *12*, 1417. <https://doi.org/10.1038/s41467-021-21696-1>.
53. Li, C.; Liu, K.; Liu, S. A Survey of Loss Functions in Deep Learning. *Mathematics* **2025**, *13*, 2417. <https://doi.org/10.3390/math13152417>.
54. Chen, S.; Guo, W. Auto-Encoders in Deep Learning—A Review with New Perspectives. *Mathematics* **2023**, *11*, 1777. <https://doi.org/10.3390/math11081777>.
55. Plusch, G.; Arsenyev-Obraztsov, S.; Kochueva, O. The Weights Reset Technique for Deep Neural Networks Implicit Regularization. *Computation* **2023**, *11*, 148. <https://doi.org/10.3390/computation11080148>.
56. Chen, L.; Yuan, C.; Xu, H.; He, Y.; Jiang, J. Robust Denoising of Structure Noise Through Dual-Diffusion Brownian Bridge Modeling and Coupled Sampling. *Electronics* **2025**, *14*, 4243. <https://doi.org/10.3390/electronics14214243>.
57. Milanfar, P.; Delbracio, M. Denoising: A powerful building block for imaging, inverse problems and machine learning. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2025**, *383*, 20240326. <https://doi.org/10.1098/rsta.2024.0326>.
58. Lu, X.; Tsao, Y.; Matsuda, S.; Hori, C. Speech enhancement based on deep denoising autoencoder. *Proc. Interspeech* **2013**, *12*, 436–440. <https://doi.org/10.21437/Interspeech.2013-130>.
59. Chen, X.; Cheng, Y.; Wang, S.; Sang, G.; Nah, K.; Wang, J. A Periodic Mapping Activation Function: Mathematical Properties and Application in Convolutional Neural Networks. *Mathematics* **2025**, *13*, 2843. <https://doi.org/10.3390/math13172843>.
60. Wenzel, M. Generative Adversarial Networks and Other Generative Models. In *Machine Learning for Brain Disorders [Internet]*; Colliot, O., Ed.; Humana: New York, NY, USA, 2023; Chapter 5. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK597493> (accessed on 13 July 2023).
61. Chechekina, O.; Tropina, E.; Fatkhutdinova, L.; Zyuzin, M.; Bogdanov, A.; Ju, Y.; Boldyrev, K. Machine learning assisted rapid approach for quantitative prediction of biochemical parameters of blood serum with FTIR spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2025**, *326*, 125283. <https://doi.org/10.1016/j.saa.2024.125283>.
62. Fadlelmoula, A.; Torres, H.; Catarino, S.O.; Minas, G.; Carvalho, V. Synthetic Data Generation Using Machine Learning Algorithms for Enhanced Type 2 Diabetes Prediction. In Proceedings of the 2025 IEEE 8th Portuguese Meeting on Bioengineering (ENBENG), Aveiro, Portugal, 25–26 September 2025; pp. 33–36.
63. Barinov, R.; Gai, V.; Kuznetsov, G.; Golubenkov, V. Automatic Evaluation of Neural Network Training Results. *Computers* **2023**, *12*, 26. <https://doi.org/10.3390/computers12020026>.
64. Protić, D.; Stanković, M.; Prodanović, R.; Vulić, I.; Stojanović, G.M.; Simić, M.; Ostojić, G.; Stankovski, S. Numerical Feature Selection and Hyperbolic Tangent Feature Scaling in Machine Learning-Based Detection of Anomalies in the Computer Network Behavior. *Electronics* **2023**, *12*, 4158. <https://doi.org/10.3390/electronics12194158>.

65. Lee, Y.-J.; Shin, N.-Y.; Lee, I.-G. Privacy-Preserving Data Sharing via PCA-Based Dimensionality Reduction in Non-IID Environments. *Electronics* **2025**, *14*, 2711. <https://doi.org/10.3390/electronics14132711>.
66. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. <https://doi.org/10.1007/s41664-018-0068-2>.
67. Nasiopoulos, D.K.; Roumeliotis, K.I.; Sakas, D.P.; Toudas, K.; Reklitis, P. Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models. *Int. J. Financ. Stud.* **2025**, *13*, 75. <https://doi.org/10.3390/ijfs13020075>.
68. Hwang, J.-S.; Lee, S.-S.; Gil, J.-W.; Lee, C.-K. Determination of Optimal Batch Size of Deep Learning Models with Time Series Data. *Sustainability* **2024**, *16*, 5936. <https://doi.org/10.3390/su16145936>.
69. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. <https://doi.org/10.1007/s42979-021-00592-x>.
70. Taha, K. Tree-based ensemble learning models for protein-protein interactions detection: A review and experimental evaluation. *BioData Min.* **2025**, *19*, 1. <https://doi.org/10.1186/s13040-025-00501-5>.
71. Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 218. <https://doi.org/10.21037/atm.2016.03.37>.
72. Kiyak, E.O.; Ghasemkhani, B.; Birant, D. High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis. *Electronics* **2023**, *12*, 3828. <https://doi.org/10.3390/electronics12183828>.
73. Salem, M.A.; Khalil, Z.A. An Interpretable Hybrid RF–ANN Early-Warning Model for Real-World Prediction of Academic Confidence and Problem-Solving Skills. *Math. Comput. Appl.* **2025**, *30*, 140. <https://doi.org/10.3390/mca30060140>.
74. Tzougas, G.; Kutzkov, K. Enhancing Logistic Regression Using Neural Networks for Classification in Actuarial Learning. *Algorithms* **2023**, *16*, 99. <https://doi.org/10.3390/a16020099>.
75. Andarabi, E.A.; Norouzian-Alam, S.; Shayganmanesh, M.; Abdolvahab, M.H. Analysis of glucose concentrations in blood solutions using FTIR and Raman spectroscopy methods. *Biomed. Opt. Express* **2025**, *16*, 2631–2662. <https://doi.org/10.1364/boe.561552>.
76. Balan, V.; Mihai, C.-T.; Cojocaru, F.-D.; Uritu, C.-M.; Dodi, G.; Botezat, D.; Gardikiotis, I. Vibrational Spectroscopy Fingerprinting in Medicine: From Molecular to Clinical Practice. *Materials* **2019**, *12*, 2884. <https://doi.org/10.3390/ma12182884>.
77. Mittal, S.; Bhargava, R. A comparison of mid-infrared spectral regions on accuracy of tissue classification. *Analyst* **2019**, *144*, 2635–2642. <https://doi.org/10.1039/c8an01782d>.
78. Baker, M.J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H.J.; Dorling, K.M.; Fielden, P.R.; Fogarty, S.W.; Fullwood, N.J.; Heys, K.A.; et al. Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **2014**, *9*, 1771–1791. <https://doi.org/10.1038/nprot.2014.110>.
79. Chiang, K.-Y.; Matsumura, F.; Yu, C.-C.; Qi, D.; Nagata, Y.; Bonn, M.; Meister, K. True Origin of Amide I Shifts Observed in Protein Spectra Obtained with Sum Frequency Generation Spectroscopy. *J. Phys. Chem. Lett.* **2023**, *14*, 4949–4954. <https://doi.org/10.1021/acs.jpcclett.3c00391>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.