

Article

Multi-Modal Deep Learning for Image Forgery Detection: A Synergistic Fusion Approach Combining Visual Artifacts and Metadata Consistency Analysis

Baysah Guwor¹  and Mohammad Shabaz^{2,*} 

¹ Faculty of Computer Applications, Marwadi University, Rajkot 360003, India; baysahguwor.128662@marwadiuniversity.ac.in

² Marwadi University Research Center, Department of Computer Engineering, Faculty of Engineering and Technology, Marwadi University, Rajkot 360003, India

* Correspondence: bhatsab4@gmail.com

Abstract

Multimedia data have been continuously increasing in magnitude, and so has the sophistication of manipulation methods, thereby making the digital forensic investigation process more complicated. The easy access to sophisticated image editing software and AI-generated materials has brought up the issue of information integrity, the reliability of legal evidence, and public trust. Traditional image forensics methods are usually concerned with either the detection of visual artifacts based on convolutional neural networks (CNNs) or based on metadata analysis, frequently independently of each other. This paper presents a multi-modal fusion paradigm, comprising visual feature-based feature extraction and metadata inconsistency-based detectors, to improve the classification strength. A two-stream design is used, comprising a high-level visual artifact capturing the transfer learning-based MobileNetV2 network and an XGBoost classifier that analyses EXIF metadata discrepancies. The heterogeneous representations are merged in a feature-level fusion strategy to generate a final authenticity prediction. It was tested on individual datasets and a compiled dataset of 26,023 images from CoMoFoD, CG-1050 and CASIA v1 and v2. The suggested approach had an overall accuracy of 83.85%, which was higher than the visual-only (68.61%) and metadata-only (75.85%) baselines. These findings show that complementary visual and metadata cues are much more useful in detection, while the use of a lightweight backbone enables efficient, high-throughput forensic analysis suitable for real-world deployment.

Keywords: digital image forensics; multi-modal learning; deep learning; Mobilenetv2; XGBoost; metadata analysis; feature fusion; multimedia security; deepfakes



Academic Editor: Michele Nappi

Received: 16 February 2026

Revised: 30 April 2026

Accepted: 9 July 2026

Published: 14 July 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.

1. Introduction

The spread of digital imaging devices, coupled with the concurrent democratization of image editing tools of great power has essentially reshaped the world of visual communication. Even though the technologies do provoke creativity, they also lower the barrier for bad actors to produce convincing forgeries. Faked images are being progressively used to launch disinformation, conduct financial fraud and produce fabricated evidence in court. As a result, digital image forensics has become a major defense mechanism, which attempts to verify the media and localize the tampering where the image is devoid of pre-embedded watermarks or signatures [1,2].

Image forgery is a more complex problem to deal with. Manipulations can take a simple form, such as copy-move, where parts of an image are cloned in order to hide objects, or a complex form, such as image splicing, where composite images are made of various sources [3], as illustrated in Figure 1. More recently, AI-based synthesis techniques have unveiled hyper-realistic fakes that can easily escape the human eye. In response to these threats, the research community has mostly converged on two different methodological paradigms; pixel-based analysis and format-based analysis [4].



Figure 1. This is an illustration of a realistic image forgery (splicing) retrieved from the CASIA Dataset. Authentic picture on the left and manipulated on the right. The red circle indicates the manipulated (spliced) region introduced into the forged image.

Pixel-based techniques take advantage of the statistical behavior of image sensors and compression algorithms. When the image is manipulated, the intrinsic statistical fingerprints of the interpolation pattern of the camera Color Filter Array (CFA), sensor noise (PRNU), or JPEG compression quantization, is usually disrupted. Convolutional Neural Networks (CNNs) and, more precisely, deep learning has been the most recent high-performing system in this area and can learn hierarchical feature representation of these subtle disturbances [3,5]. Visual forensic tools are, however, susceptible to counter-forensic activities like resizing, re-compression or adversarial blurring, which can erase the delicate high/frequency artifacts that are the signs of compromised forensic evidence.

On the other hand, metadata-based forensics looks at those supplementary data inscribed in image files like Exchangeable Image File Format (Exif) tags [6]. These tags have an important history in capturing the device, exposure settings and software used in processing. Metadata inconsistencies, such as an incompatibility between quantization tables specified in the Exif header and the statistics of the actual image, or the existence of editing software signatures, are good clues to manipulation. However, metadata can be unreliable when used in isolation; it can be deprived, changed, or even just lost in transit across social media platforms, and thus this modality is not useful on its own.

It is proposed in this research that such unimodal limits can be overcome by using a multi-modal fusion approach with a strong forgery detection system. We claim that complementary evidence is visual artifacts and metadata inconsistencies. A visually convincing forgery might betray itself through anomalous shutter-speed tags, while a stripped-metadata image might still exhibit tell-tale splicing boundaries in the pixel domain. Consider a scenario where a malicious actor splices a person into a new background.

- ✓ To fool the eye, they can meticulously blur the edges and color-correct the splice to cheat the eye.

- ✓ They may, however, not update the Exif Make and Model tags, or the software may leave a signature such as Adobe Photoshop 24.0 in the Software tag.
- ✓ On the other hand, an attacker can strip all metadata in order to conceal their trails. The anomalies in the visual noises at the splice boundary in this instance emerge as the main evidence.

In our effort to fill these gaps, we introduce a unified, scalable and robust multi-modal forensic framework. We make the following contributions:

- ✓ We construct a parallel processing pipeline in which a MobileNetV2 backbone is used to extract deep visual features and an XGBoost model is used to extract structured metadata attributes. This composite strategy uses the capability of feature-learning of deep learning with unstructured pixel data and the efficiency of gradient boosting with structured metadata.
- ✓ We investigate a fusion mechanism that concatenates high-level feature representations from the two streams, enabling the model to learn an effective weighting between visual and metadata evidence.
- ✓ To deal with the problem of data fragmentation in forensics research, we aggregate four benchmark data sets: CASIA (v1 and v2), CoMoFoD and CG-1050. This provides a total of 26,023 original and manipulated images, all of different resolution, compression, and manipulation (splicing, copy-move, retouching).
- ✓ Finally, we also provide a general analysis which demonstrates that the fusion model has statistically significantly higher detection accuracy and AUC-ROC than each of the visual and metadata models individually.

As for the rest of the paper, it is structured as follows: Section 2 includes the review of the related work. Section 3 reflects our approach to the methodology, which is the architecture design and fusion strategy. Section 4 describes our experiment background and data. Section 5 presents the results and discussion, comparative analysis of the baseline methods, limitations and implications. Finally, we have a conclusion and future research directions in Section 6.

2. Related Work

With the intersection of deep learning and digital forensics, the image forgery-detection landscape has changed greatly. Moreover, with the spread of generative AI and advanced editing software, the problem of digital image authenticity has increased and is becoming more alarming. In this section, we discuss recent works that have been carried out. The literature is divided into three main areas, which include the detection of visual artifacts, metadata forensics, and new multi-modal fusion models.

2.1. Visual Artifact Detection

The initial approaches to visual forensics were focused on the detection of certain types of evidence of manipulation with the help of signal processing mechanisms. The analysis of the JPEG compression artifacts is examined through Error Level Analysis (ELA), and the discontinuity in sensor Bayer patterns can be detected as Color Filter Array (CFA) interpolation analysis. Since the development of deep learning, Convolutional Neural Networks (CNNs) have shown better learning ability of discriminative features in areas of manipulation.

A deep learning model for image forgery detection with inductive transfer learning was put forward by Bevinamarad et al. [7], and in this framework, to detect forgery, the network is not trained, but fine-tuned, using the previously trained feature maps. Systematic hyperparameter optimization was used to evaluate the model on CUISD and

CoMoFoD datasets. It registered 89.0 percent precision, 86.43 percent recall, and 87.32 percent F1-score, as well as 87.72 percent accuracy.

Weng et al. [8] suggested UCM-Net, a U-Net-based architecture that detects copy-move forgery through a differentiable approach to large and small tampering regions, with the support of deep and lightweight backbones. It combines asymmetric across-layer connections, self-correlation, atrous spatial pyramid pooling, and multi-depth residual U-blocks, to capture multi-scale contextual features.

Yan et al. [9] came up with DIF-Net, which is a two-stream intermediate fusion network that has pyramids of adaptive convolutional depth-wise strip convolutions that extract multi-scale features. The learnable fusion modules are trained to weight features in streams dynamically at a level that captures channel and spatial consistency to enhance the quality of the mask. The technique achieved 3.3 percent F1 and 2.4 percent AUC improvements in comparison with the previous methods on CASIA2. In a conference paper, a CNN-based copy-move detector system [10] used block-wise transform features with serial convolutional layers, having 97.7% classification rate on CASIA2, and asserted that it could process in real-time.

In spite of the great progress, there are some restrictions regarding visual-only methods. They are susceptible to adversarial isolations, have a hard time with highly compressed photos where high-frequency forensic evidence is weakened, and can fail to generalize across datasets because they are overfitting to dataset-specific artifacts. Moreover, such methods are not able to identify metadata manipulation or authentic photos, which are visually clean and contextually fabricated.

2.2. Metadata Forensics

As the visual analysis is concerned with pixel-level artifacts, metadata forensics is concerned with the structural and semantic integrity of file headers. The Exchangeable Image File Format (Exif) standard contains critical capture parameters such as Make, Model, Date-Time-Original, Software, GPS-Info and JPEG Quantization Table (DQT). The inconsistencies in these areas can demonstrate post-capture manipulation, since editing software tends to leave characteristic traces; the Adobe Photoshop application, for example, tends to add its identifier in the Software tag and changes the Modify-Date field.

Metadata-only-based methods of detecting image forgery are based on file-related characteristics and not on pixel information, and have important computational benefits. Bedi et al. [11] suggested a neural network to identify the features of a camera, such as camera make, model, and software version, in JPEG metadata, to detect an authentic and a manipulated image. Manipulation operations that are identified in their model are cropping, rotation and compression. The compact input needs of the system rendered it more efficient than the pixel-based alternatives, but were mostly only validated in controlled JPEG environments [11].

Another study named Exif2Vec, by Umair et al. [12], was based on the pretrained Word2Vec embeddings, which they used to identify semantic inconsistencies in tags such as GPS and timestamp. The framework determines the metadata values that appear to have no logical relation to the common capture patterns by mapping tags into the vector space. The method was evaluated on fact-checked datasets and attained up to 80 percent accuracy in the absence of pixel information. This technique is efficient at detecting indications of tampering in crowd-sourced images, but performance is still mediocre when compared to hybrid methods that use the visual features.

In turn, Lee et al. [13] came up with a forensic approach that combines EXIF parsing, a discrete quantization table (DQT), analysis, and reference database matching in order to identify traces of manipulation. This technique increased explainability by identifying

particular tags that had changed, but it needs to keep databases of camera-specific signatures. Self-supervised learning is a new trend, and Zhong et al. [14] suggest that camera metadata-based detection could be utilized in images generated by AI in label-saving conditions.

Regarding its potential, metadata forensics is seen to be discounted in the literature because of its fragile nature. Exif data is regularly removed on social media sites (Facebook, Instagram, and Twitter) to reduce file size and safeguard user privacy. Metadata can be lost accidentally by the casual user via format conversion, or by taking a screenshot. Furthermore, advanced adversaries can easily forge or transplant metadata on images and make metadata-only methods unreliable. This leads to the fact that very few studies are trying to formalize metadata and make it a part of the end-to-end deep learning pipeline.

2.3. Multi-Modal Fusion

Recent studies recommend the fusion of the heterogeneous sources of information, in order to overcome the constraints of single-modality. The reason is quite convincing: the analysis of consistency between what the picture looks like (visual artifacts) and what the file claims to be (metadata provenance).

Xia et al. [15] have proposed DMFF-Net, a double-stream multilevel fusion architecture, a model that explicitly integrates RGB and noise features at primary, intermediate and advanced levels of hierarchy. Graded fusion is done by the Primary Feature Fusion Module (PFFM) and Advanced Feature Fusion Module (AFFM), with multi-supervision decoding producing boundary, regular and refined segmentation-masks. Cross dataset validation experiments on CASIA, Columbia, COVERAGE, NIST16 and the real-world IMD20 data showed good localization performance. The multi-modal designation, however, is only applied in the situation of streams in the visual domain (RGB vs. noise residuals); the framework is not designed to carry non-visual metadata streams.

He et al. [16] came up with MMFA-Net, which uses multi-head fuse attention to accomplish localization and detection by fusing multi-modal visual features, both spatially and channel-wise. In another study, He et al. [16] designed a spatial-frequency fusion network, which is based on face-forgery detection. A small U-Net produces pixel-resolved saliency maps which are used to fuse spatial and frequency features (extracted using a more effective Swin Transformer and Haar wavelet transform, respectively). Complementarity is perfected by cross-domain attention, channel recalibration, and spatial gating modules. The method was also evaluated on FaceForensics++ and Celeb-DFv2.

Gong et al. [17] used the Cross-Attention Fusion Module (CAFM) and Guided Attention Module (GAM) to integrate both spatial and frequency domain features, and a Multi-Scale Feature Fusion Module (MSFFM) to obtain hierarchical artifacts. The framework was reported to achieve 3.9% average AUC improvement compared to GocNet on FaceForensics++ and WildDeepfake benchmarks.

The critical gap for the present survey is that of multi-modal approaches, most of which primarily combine features of the visual domain, but do not include the integration of symbolic metadata-feature forensics. Combining high-dimensional CNN embeddings with structured Exif feature vectors is still under-investigated. It is this particular gap that our work is directed at, as we offer a mechanism to combine the visual artifact detection and metadata consistency analysis into a unified deep learning framework. We have summarized the related work in Table 1.

Table 1. Summary analysis of related work.

Study	Core Methodology	Datasets	Modality		
			Multi-Model/ Hybrid	Visual	Metadata
Bevinamarad et al. [7]	Deep learning (DL) framework (with inductive transfer learning (ITL))	CoMoFoD and CUISD		✓	
Weng et al. [8]	U-Net with asymmetric cross-layer connections and self-correlation	CASIA, CoMoFoD, COVERAGE		✓	
Yan et al. [9]	Dual-stream intermediate fusion with adaptive convolutional pyramid	CASIA2	✓	✓	
Lee et al. [13]	Forensic parsing of EXIF, DQT, filename vs. Reference DB	Custom manipulated dataset			✓
Xia et al. [15]	Double-stream multilevel RGB-noise fusion (DMFF-Net)	CASIA, Columbia, COVERAGE, NIST16, IMD20	✓	✓	
Umair et al. [18]	Exif-BERT integration in context-aware framework	-			✓
He et al. [16]	Saliency-guided spatial–frequency fusion with Swin Transformer	FaceForensics++, Celeb-DFv2	✓	✓	
Gong et al. [17]	Cross-attention fusion of spatial–frequency with multi-scale integration	FaceForensics++, WildDeepfake	✓	✓	
Liu and Yao [19]	Multi-head fuse attention for spatial-channel fusion	-	✓	✓	
Proposed Framework	MobileNet-V2 (visual) + XGBoost (Exif metadata) with feature-level fusion mechanism	CASIA-v1, CASIA-v2, CoMoFoD, CG-1050	✓	✓	✓

Note: (✓) indicates that the corresponding modality is used/supported by the study. While (-) indicates that the information is not reported or not applicable.

3. Methodology

In this section, the proposed multi-mode fusion framework is described. The system is designed such that it classifies an input image I into one of two classes: $y \in \{0, 1\}$, where 0 is Authentic, and 1 is Manipulated. The framework works with two parallel streams on the input image: the Visual Stream (S_{vis}), which derives spatial forensic features, and the Metadata Stream (S_{meta}), which extracts file header consistency. In Figure 2, the general methodology flowchart is presented.

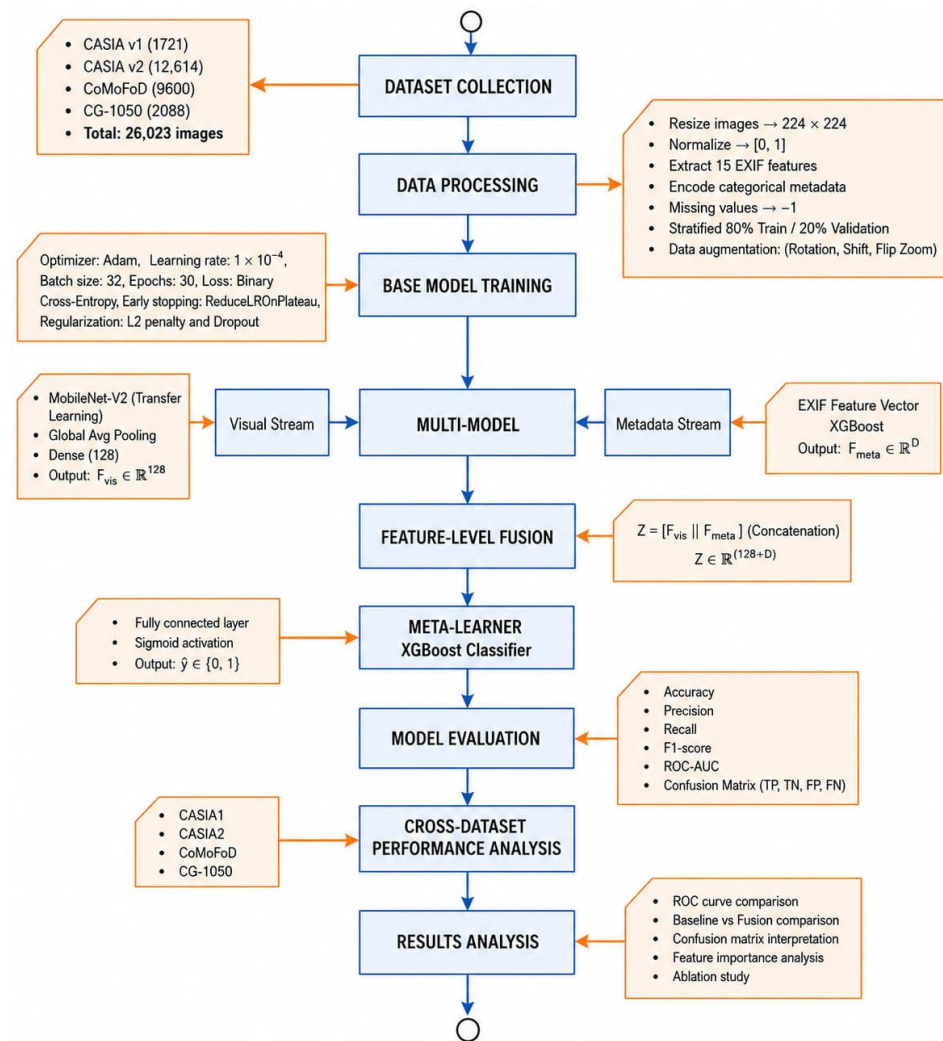


Figure 2. Methodology flowchart for proposed approach.

3.1. MobileNet-V2 with Transfer Learning

The visual stream tries to capture manipulation artifacts which remain in the spatial representation, e.g., inconsistencies of edges in splicing areas or statistical anomalies in smoothed edges.

Architecture Selection: a variety of convolutional neural network architectures may be used to extract visual artifacts, but this paper uses MobileNet-V2 [20] because of its favorable trade-off between performance and computational efficiency. It was empirically compared to well-known architectures, such as VGG-16, ResNet-50, and EfficientNet-B4. Within a consistent experimental setup, the heavier models obtained similar baseline performance at the cost of much greater inference latency, with an average of 92–96 ms per image.

In contrast, MobileNetV2 has many fewer requirements and the highest inference time of only 81.40 ms per image, while maintaining competitive feature-representation capability. This efficiency and performance balance makes MobileNetV2 a good fit for the real-life forensic uses, where it is necessary to process large volumes of images as quickly as possible. Moreover, it has a lightweight architecture enabling the computational resources to be efficiently assigned to the metadata analysis step in the proposed fusion framework.

It presents the Inverted Residual with Linear Bottleneck block that is especially useful in storing information in low-dimensional manifolds. The basic unit is the Depth-Wise

Separable Convolution, which splits a conventional convolution into two separate layers, as follows [21]:

- ✓ Depth-wise Convolution: it applies a single filter per input channel, and
- ✓ Pointwise Convolution: 1×1 convolution that combines the outputs of the depth-wise layer.

Mathematically, for an input tensor of size $h \times w \times d_{\text{in}}$, a standard convolution kernel of size $k \times k \times d_{\text{in}} \times d_{\text{out}}$ has a computational cost of the following: $\text{Cost}_{\text{std}} = h \cdot w \cdot d_{\text{in}} \cdot d_{\text{out}} \cdot k^2$. On the other hand, the depth-wise separable convolution cost is $\text{Cost}_{\text{dw}} = h \cdot w \cdot d_{\text{in}} \cdot (k^2 + d_{\text{out}})$. The vanishing gradient penalties are generally removed in factorization, which results in computation time that is a factor of 8-to-9 smaller, enabling much deeper networks.

Feature extraction and embedding: the backbone (MobileNet) is already pre-trained on the ImageNet dataset to initialize the weights with robust low-level feature detectors (edges and textures) [22]. To maintain these generic characteristics and to avoid overfitting to the smaller forensic dataset, we freeze the early layers.

Let $f_{\theta}(I)$ denote the mapping of the frozen MobileNetV2 backbone. The result of this network is a high-level feature map tensor $T \in \mathbb{R}^{7 \times 7 \times 1280}$, whose spatial dimensions can identify localized visual patterns and whose channel dimension can identify responses to semantic features. In order to reduce this tensor to a small and discriminative representation, we add a discriminative classification head tailored to forensic-feature aggregation.

1. Global Average Pooling (GAP): this operation summarizes each feature channel by computing its spatial average, effectively reducing sensitivity to exact object location, while preserving semantic information:

$$v_{\text{gap}} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W T_{i,j}.$$

The resulting vector $v_{\text{gap}} \in \mathbb{R}^{1280}$ provides a global description of the image content.

2. Dense Projection: a fully connected layer then projects this high-dimensional vector into a compact embedding space that is more suitable for fusion and classification:

$$F_{\text{vis}} = \sigma(W_1 \cdot v_{\text{gap}} + b_1)$$

where $W_1 \in \mathbb{R}^{128 \times 1280}$ and σ denotes the ReLU activation function. The resulting embedding $F_{\text{vis}} \in \mathbb{R}^{128}$ represents the final visual-feature vector used in the multi-modal fusion stage.

Visual Loss Function: during training, the visual branch is optimized to distinguish between authentic and manipulated images by minimizing the binary cross-entropy loss:

$$L_{\text{vis}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_{\text{vis},i}) + (1 - y_i) \log(1 - \hat{y}_{\text{vis},i})]$$

This loss measures the discrepancy between the ground truth labels y_i and that of the predicted probabilities $\hat{y}_{\text{vis},i}$, where $\hat{y}_{\text{vis}}$ is obtained from a sigmoid-activated auxiliary classifier attached to F_{vis} during the pre-training phase. This auxiliary supervision helps the visual embedding learn discriminative forensic features before fusion.

3.2. XGBoost Classifier

The metadata branch is concerned with the structured data that is present in the header of image files and, in many cases, it can contain inconsistencies added in the course of editing or re-saving an image. We use XGBoost (Extreme Gradient Boosting) [23], which is

a scalable and robust tree-based learning framework that is most suitable to heterogeneous data and partially missing data.

Feature Engineering and Encoding Raw Exif metadata is mostly categorical, and often not complete, particularly in manipulated images. To deal with this, we model a piping of preprocessing:

- ✓ Tag Extraction: a set of K informative Exif tags is extracted. Amongst these are included Make, Model, Software, Modify-Date, Date-Time-Original, F-Number, ISO and Exposure-Program, among others. These tags record the details of the camera hardware, capture settings, and post-processing software.
- ✓ Consistency Checks: the logical relationships among the tags are applied in order to generate the binary indicators of possible manipulation:

$$x_{\text{conflict}} = \mathbb{I}(\text{ModifyDate} > \text{DateTimeOriginal} + \epsilon)$$

This functionality brings out suspicious temporal discrepancies, whereas $x_{\text{editing}} = \mathbb{I}((\text{"Adobe"} \vee \text{"GIMP"}) \in \text{Software})$ clearly indicates the use of image-editing software.

- ✓ Encoding: categorical tags like camera make are label-encoded in a manner that permits tree-based learning. Stripped or sanitized images simply contain missing values, encoded into a particular type of integer (e.g., -1), and the absence of metadata is a predictive variable that can be used in the model.

Gradient-Boosted Decision Trees: XGBoost generates an ensemble of K regression trees $\{t_1, t_2, \dots, t_K\}$, each which represents different patterns in the metadata space. The ultimate prediction value of a given metadata feature vector x_m is calculated as the sum of the outputs of all trees:

$$\hat{y}_{\text{meta}} = \sum_{k=1}^K t_k(x_m)$$

The model is trained iteratively. At each of the training steps, t , a new tree, t_t , is added, to minimize the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^N l(y_i, \hat{y}_i^{(t-1)} + t_t(x_i)) + \Omega(t_t)$$

where l denotes the log-loss function, and $\Omega(t) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ acts as a regularization term that penalizes overly complex trees. This regularization is essential for preventing the model from memorizing camera-specific identifiers or dataset-specific artifacts.

Table 2 is a summary of the metadata features that were extracted in the feature engineering step. PIL.ExifTags is used in the extraction in conjunction with native file operations to extract 14 standard EXIF attributes and a structural feature (FileSize), which creates a 15-dimensional vector utilized by the XGBoost classifier.

Table 2. Extracted metadata and Exif features.

No.	Name of Feature	Description/Forensic Relevance
1	Make	Manufacturer of the recording device
2	Model	Device model name or number
3	Software	Software/firmware used (e.g., Photoshop indicators)
4	DateTime	Creation or last modification timestamp
5	ExifImageWidth	Image width in pixels

Table 2. Cont.

No.	Name of Feature	Description/Forensic Relevance
6	ExifImageHeight	Image height in pixels
7	ExposureProgram	Camera exposure mode
8	ISOSpeedRatings	Sensor sensitivity (ISO)
9	ShutterSpeedValue	Shutter speed at capture
10	ApertureValue	Lens aperture at capture
11	MeteringMode	Light metering method
12	Flash	Flash usage indicator
13	FocalLength	Lens focal length (mm)
14	ColorSpace	Color space (e.g., sRGB, Adobe RGB)
15	FileSize	File size in bytes (structural indicator)

3.3. Multi-Modal Fusion Strategy

The visual and metadata information is difficult to combine because of the dissimilarity in the feature dimensionality (128 vs. about 15) and the data type (continuous vs. discrete). As a solution, we adopt a feature-level fusion approach that combines the intermediate representations of both modalities, namely the 128-dimensional MobileNetV2 embedding and the 15-dimensional EXIF feature-vector, and that trains a unified classifier on this space. This makes sure that the model learns interaction between cross-modal features, and not just a combination of two independent predictions.

Let $F_{vis} \in \mathbb{R}^{128}$ denote the visual embedding and $F_{meta} \in \mathbb{R}^M$ denote the metadata representation. These vectors are concatenated to form a joint feature vector:

$$Z = F_{vis} \parallel F_{meta}$$

The fused representation Z is then passed to an XGBoost meta-classifier, which learns decision boundaries through an ensemble of gradient-boosted trees to produce the final prediction $\hat{y}_{final}$. This fusion design enables adaptive reliance between modalities, depending on the available signal:

- Case A (Stripped Metadata): when metadata is absent or unreliable, the model primarily relies on visual cues encoded in F_{vis} .
- Case B (High-Quality Deepfake): when visual artifacts are subtle, strong metadata indicators (e.g., Software = “Adobe Photoshop”) can guide the model toward detecting manipulation.

4. Experimental Setup

In order to evaluate the multi-modal fusion framework, we developed a multi-faceted experimental protocol that includes a variety of datasets. In this section, our methodology for the experiment is described.

4.1. Datasets and Preprocessing

In order to guarantee the strength and generalizability of our framework, we have created a coherent unified forensic dataset by summing 4 popular benchmarks: CASIA (v1 and v2), CoMoFoD, and CG-1050 v2. This aggregation creates critical diversity of image resolution, file format, compression quality, and methods of manipulation.

- ✓ CASIA Dataset: the CASIA data set is the most mentioned in the forgery-detection studies. CASIA v1 contains 1721 images (800 original, 921 manipulated). The manipulated images mostly utilize splicing and copy-move attacks. CASIA v2 is a much larger dataset of 12,614 images (7491 authentic, 5123 manipulated). This version has more accurate forgeries with different lighting effects and a difference in scale.
- ✓ CoMoFoD Dataset: the CoMoFoD Dataset is designed to detect copy-move (cloning) attacks, in particular. It includes 9600 photos (4800 authentic, 4800 manipulated). The manipulated images are built by copying a portion of an image and pasting it within another location of that very same image. The dataset is also supplemented by rotation of the copied area, scaling transformations, and JPEG post-compression at different quality levels. This renders CoMoFoD especially convenient for testing strong performance with regard to geometric transformations.
- ✓ CG-1050 v2 Dataset: the CG-1050 v2 dataset is designed for digital image forgery-detection research. It contains 1044 authentic images and 1044 manipulated images generated using manipulations such as copy-move, cut-paste, retouching, and color adjustments. The dataset also provides benchmarking and evaluation of image-manipulation detection and localization algorithms in multimedia forensics.
- ✓ Merged Dataset: Table 3 presents the statistics of the classes within the unified dataset. In order to avoid the bias of models representing over-represented classes, we also used stratified sampling method in order to divide the data into 80/20 training-and-validation sets, so that the proportion of authentic to manipulated images is equal in both subsets.

Table 3. Composition of the unified forensic dataset.

Dataset	Types of Manipulation	Image Formats	Authentic (0)	Manipulated (1)	Total Images
CASIA v1 [24]	copy-move, splicing	JPG	800	921	1721
CASIA v2 [24]	copy-move, splicing	JPG, BMP, TIF	7491	5123	12,614
CoMoFoD [25]	copy-move forgery	JPG, PNG	4800	4800	9600
CG-1050(v2) [26]	copy-move, cut-paste, retouching, and colorizing	JPG	1044	1044	2088
TOTAL			14,135	11,888	26,023

4.2. Implementation Details

The experiment was carried out using a high-performance computing platform that was powered by an NVIDIA Tesla P100 GPU and a 16GB system RAM. The software stack consists of the following: TensorFlow 2.19, to build deep learning models and train them; XGBoost 3.1.0, for gradient-boosting structured metadata; and Python 3.11, for using additional packages including NumPy, Pandas, Scikit-learn, PIL, and ExifRead, to extract metadata. To train the visual branch, the following protocol is used:

- ✓ Preprocessing: resize images to 224×224 pixels and normalize them to either 0 or 1.
- ✓ Data Augmentation: to combat overfitting, we apply random augmentation during training: Rotation: 20° , Width/Height Shift: 20% of image dimensions, Horizontal Flip: 50% probability, Zoom: 10% range
- ✓ Optimization: the Adam optimizer is used, along with initial learning rate $= 1 \times 10^{-4}$.
- ✓ Learning rate schedule: “ReduceLROnPlateau” callback reduces α by a factor of 0.5 when validation loss plateaus for 3 consecutive epochs.
- ✓ Early Stopping: the training process halts if the validation AuC does not improve for 5 successive epochs, restoring the best weights.
- ✓ Batch Size: 32 images per batch.

- ✓ Epochs: 30 epochs.
- Training protocol for the metadata branch includes
- ✓ Feature set: fifteen (15) Exif fields are extracted per image.
- ✓ Missing data handling: label-encoded as “-1” and treated as informative.
- ✓ XGBoost hyperparameters: max-depth = 6, learning-rate = 0.1, n-estimators = 100, colsample_bytree = 0.8, subsample = 0.8, objective = binary:logistic

4.3. Evaluation Metrics

In order to evaluate the performance, we have use standard binary classification metrics as described below. For a binary classifier with predictions $\hat{y}$ and ground truth y :

1. Accuracy:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision:

$$Prec = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity):

$$Rec = \frac{TP}{TP + FN}$$

4. F1-Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. AUC-ROC: Area under the Receiver Operating Characteristic curve is used to plot the (True Positive Rate) versus (False Positive Rate) across all levels of the threshold values.

where

- ✓ TP: True Positives (where the model correctly identified manipulated images).
- ✓ TN: True Negatives (here, the model correctly identified authentic images).
- ✓ FP: False Positives (for this, the model flagged authentic images as being manipulated).
- ✓ FN: False Negatives (the model flagged manipulated images as authentic).

5. Discussion of the Results

In this section, the detailed experimental results will be presented through an evaluation of the baseline models and the hybrid framework. The results are compared and analyzed in terms of performance measures, such as accuracy, precision, recall, F1-score, and ROC-AUC scores.

5.1. Individual- and Merged-Dataset Performance Analysis

For evaluating the proposed model, we cut across the four individual datasets as mentioned and as indicated in Table 3. The entire assessments were made under the same parameters as explained above, so as to achieve a balance. All the datasets were divided into 80 percent training and 20 percent validation. This discussion shows that the model is flexible to various types of forgery and properties of source devices. The highest accuracy of the model is 99.46% in the CASIA2 dataset, which shows really impressive sensitivity for advanced forensic patterns. The summary of the results is shown in Figure 3.

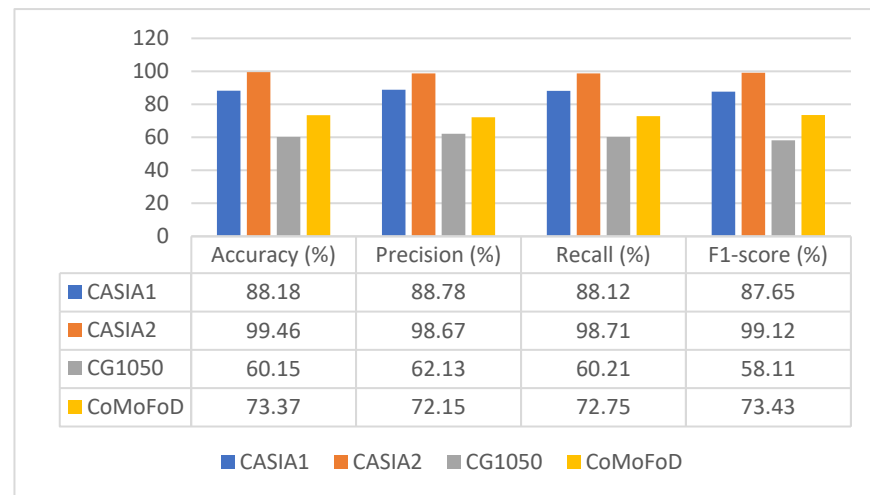


Figure 3. The performance indicators of the suggested model across individual datasets.

With the fusion method, we achieved 88.18% accuracy on the CASIA-1 dataset. Splicing is one of the minor manipulations that are included in this dataset. The framework achieved high precision of 88.78% and recall of 88.12%, which implies that the combination of XGBoost metadata analysis and MobileNet visual features are useful in identifying forgery artifacts. The CASIA2 forensics dataset had an impressive accuracy level of 99.46%. It is a very high performance level, which implies that the model was indeed able to learn the intricate patterns of advanced manipulations that are contained within this dataset. The high F1-score also confirms the strength of the fusion method in the case of high-quality manipulated evidence. The framework attains approximately 60.15% accuracy on the CG-1050 v2 dataset, way below its accuracy on full-image photographic challenges such as CASIA. Such a performance gap is an immediate outcome of the particular composition of the dataset. CG-1050 v2 is composed of localized cropped blocks (patches) and not entire images, as described by the authors of the datasets [26]. This is a limitation that severely affects both of our modalities. First, the automated cropping scripts which create such blocks, naturally strip original EXIF headers, which makes our metadata stream uninformative for this group. Second, evaluating isolated patches denies the MobileNetV2 visual-stream global contextual information, including lighting anomalies and background noise variations, which it counts on to determine forgery boundaries.

The CoMoFoD dataset provided an accuracy of 73.37%. This data is a difficult set of copy-move forgeries. The balanced measurements between precision and recall prove that the adaptive fusion mechanism is effective in balancing the visual and metadata streams.

To evaluate cross-dataset generalization, we trained and validated information on the combined dataset, comprising equal samples of all four sources. This becomes important in contemporary forensic research, since single-dataset evaluation tends to exaggerate or understate the effectiveness of a model in the real world as a result of dataset bias. The combined dataset contains a total of 26,023 image files that were divided into an 80-percent training set of 20,817 and a 20-percent validation set of 5206, respectively, using a stratified data-splitting method. The accuracy, precision, recall and f1-score of the merged dataset experiment are approximately 83.85%, 84.78%, 80.12%, and 82.33%, respectively, on the proposed model (see Figure 4). These encouraging findings indicate that the framework is able to support heterogeneous evidence types under a single unified model, which can be used to remove the necessity of different types of forensic cases being supported using specialized models. A more detailed analysis is discussed in Section 5.2 and sub-sections.

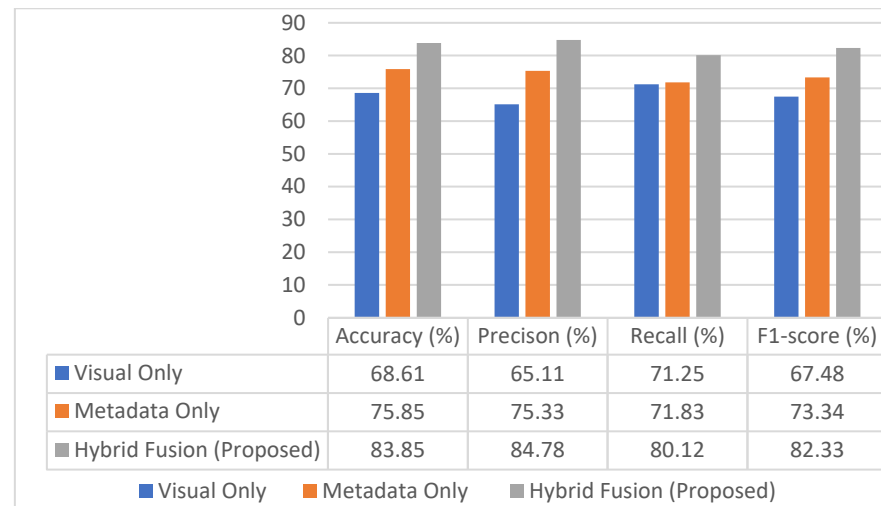


Figure 4. Performance metrics of baseline and hybrid models on the merged dataset.

5.2. Proposed Hybrid-Model Performance Evaluation and Comparison Baseline Models

In order to create a good comparative base on which the proposed hybrid architecture would be assessed, the individual pathways were tested individually on the combined data. This dataset, which contained a balanced number of authentic and manipulated images of all four sources, was a good yardstick to evaluate the ability of each modality.

MobileNet-V2, the visual-only baseline, had an accuracy of 68.61% on validation, a precision of about 65.11%, and a recall of 71.25%. Although MobileNetV2 is useful in the identification of spatial and textual artifacts, its efficiency appears to be low in situations where manipulations cause only slight visual effects or get lost in the compression. This can also be attributed to the fact that our main focus lies in the performance of the fusion model. XGBoost, which is the metadata-only baseline, gave a performance accuracy of 75.85%. This indicates that, in this particular arrangement of the dataset, structured metadata (EXIF tags, file size, and device attributes) contains a more powerful forensic cue than visual features.

The feature-level fusion model was evidently better than both baselines, with an accuracy of 83.85%, precision 84.78%, recall, 80.12% and F1-score 82.33%. This finding highlights the effect of the feature-level fusion mechanism. The framework implicitly learns to weight both pathways through tree-based feature splits, which means it pays more attention to the cues that will yield better results in a particular image, which improves its capability of discriminating between an original and a manipulated image.

Both the training and the validation accuracies of the MobileNet model are illustrated in Figure 5. The training curves indicate steady improvement throughout the epochs with little overfitting, but a minor shift has been noted in both the training and validation loss after epoch 15, potentially because of heightened sensitivity to finer details of generative features. Figure 6 shows the normalized feature-importance analysis of the XGBoost metadata model with ColorSpace, Software, and FileSize emerging as the most influential attributes. From a forensic perspective, certain values of these features can be indicative of manipulation. As an example, identifiers like Adobe Photoshop or Gimp are seen in manipulated pictures, replacing original camera firmware records. Likewise, ColorSpace transitions (e.g., to sRGB) imply re-encoding in post-processing, whereas abnormal FileSize values indicate recompression or structural changes.

These trends suggest that the model uses significant metadata inconsistencies that are related to the image-editing pipeline.

In order to visualize the comparative performance, ROC curves and confusion matrices of the basic models and the proposed hybrid model are compared.

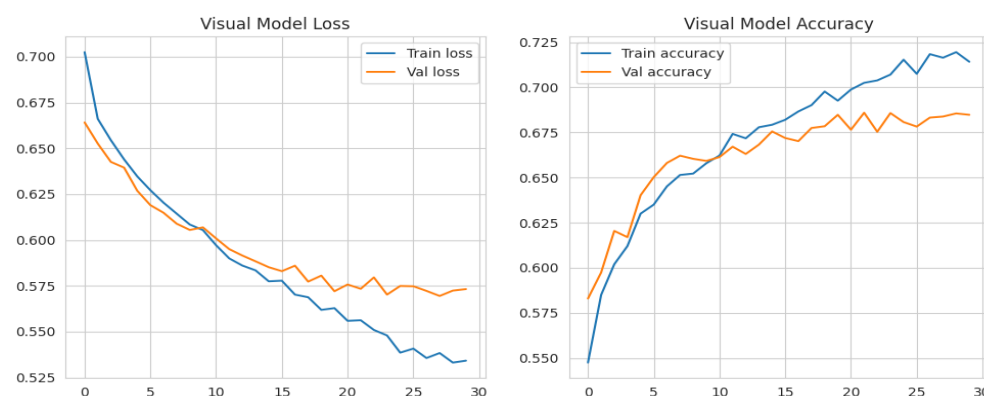


Figure 5. Training and validation trends for the visual model across 30 epochs.

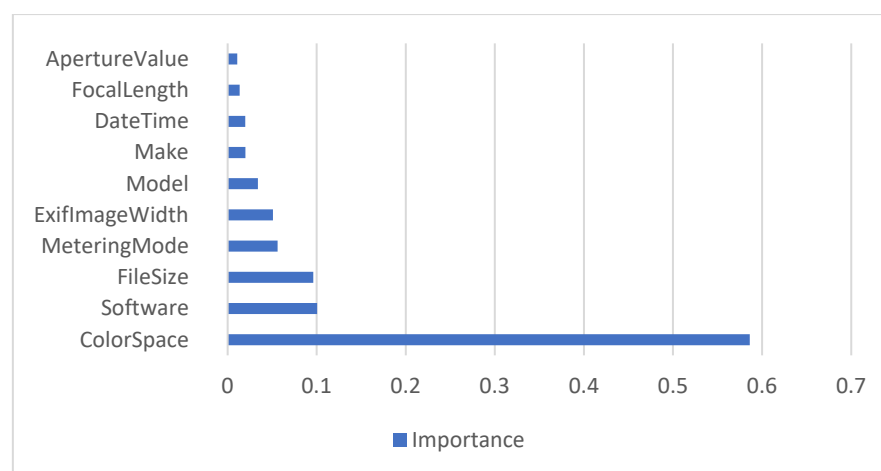


Figure 6. Feature importance extracted from the XGBoost metadata model.

The comparative ROC curves of the hybrid and baseline models are shown in Figure 7. The proposed model has an area under the curve (AUC) of 0.92, and this shows that the model has better classification performance as compared to each independent baseline model. Its ROC curve stays steadily higher than the baselines, especially in the low false-positive area, and this once again confirms its increased sensitivity and specificity in the detection of manipulated images.

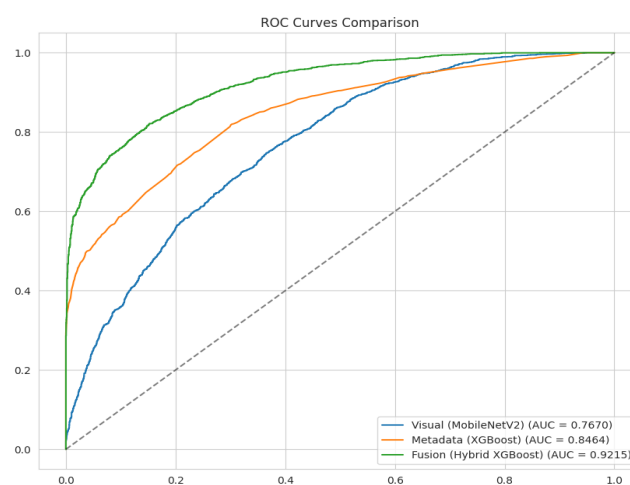


Figure 7. ROC Curve Comparison (baselines vs. fusion model).

In order to offer a comparative diagnostic analysis of classification behavior, Figures 8 and 9 illustrate a comparison of the confusion matrix, including the MobileNetV2 visual model, the XGBoost metadata model, and the proposed hybrid model. Class 0 in these matrices is the authentic images and class 1 is the manipulated images. The matrices give in-depth information regarding the distribution of true negatives, true positives, false positives, and false negatives.

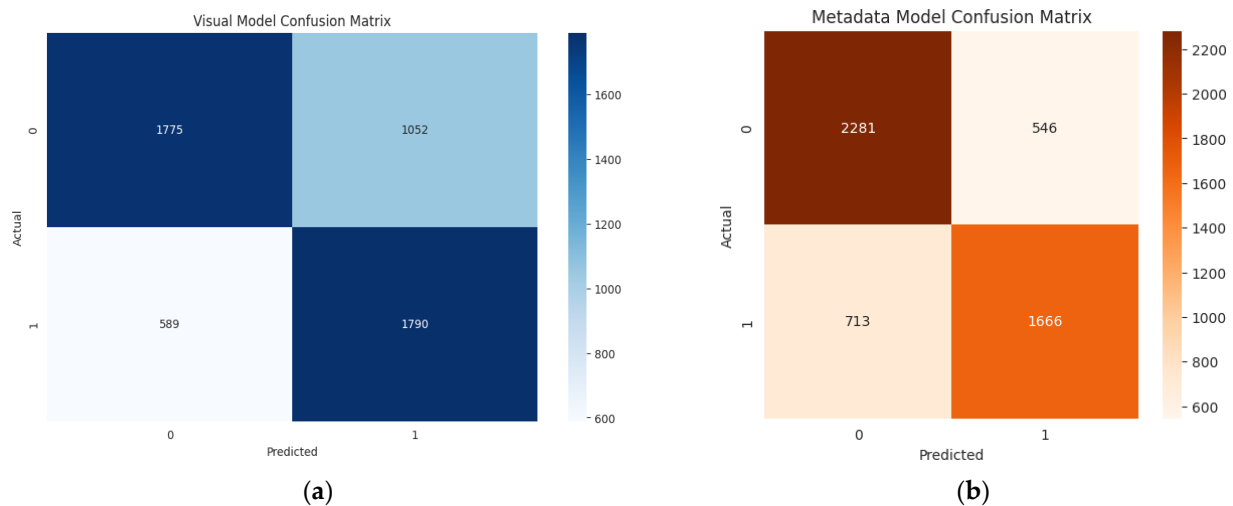


Figure 8. Confusion matrices of the baseline models evaluated on the combined dataset: (a) visual baseline model(MobileNetV2); (b) metadata baseline model (XGBoost).

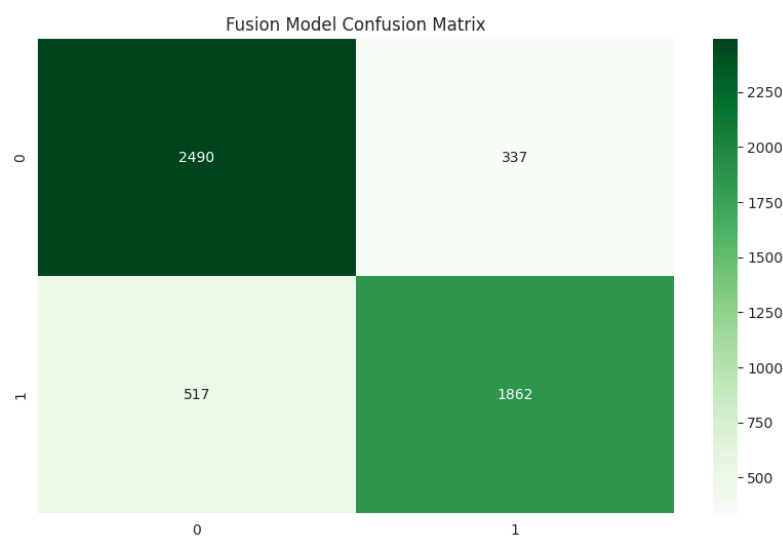


Figure 9. Confusion matrices of the proposed fusion model.

In the case of the visual-only MobileNetV2 model, 1775 authentic images were correctly classified as authentic (true negatives), and 1790 manipulated images were correctly classified as manipulated (true positives). Nevertheless, 1052 authentic images were falsely classified as manipulated, and 589 manipulated images were falsely classified as authentic. These findings suggest that the visual model identifies a significant number of manipulated images, but generates a relatively large false alarm.

In the XGBoost model, 2281 authentic images were correctly classified, which means true negatives, and 1666 manipulated images were also correctly classified as true positives. There were 546 false positives and 713 false negatives developed by the model. The metadata pathway is much more discriminative than the visual model in terms of false positives, which proves that structured EXIF information is a powerful discriminative feature in determining authentic and manipulated content.

The feature-level fusion model proposed had the best and most balanced classification performance, as seen in Figure 9. It has properly identified 2490 genuine images (true negatives) and 1862 manipulated images (true positives), and minimized the number of false positives, to 337, and also false negatives, to 517. The decrease in the two types of errors proves the efficiency of the feature-level fusion mechanism. The fusion method enhances reliability and strength in manipulated- and authentic-image identification through integrating complementary visual and metadata capabilities.

5.2.1. Stability and Training Efficiency

The suggested model showed consistent behavior with respect to training, converged in 30 epochs. On average, it took about 418 s to run on a standard GPU hardware (Tesla P100) to complete each epoch of training. This efficiency facilitates implementation in operational forensic laboratories.

5.2.2. Modality Contribution and Adaptive Behavior

To analyze the contribution of each modality within the feature-level fusion framework, we examined the feature-importance scores derived from the trained XGBoost classifier. The accumulated importance explains that the 128-dimensional visual embedding accounts for approximately 78.4% of the model's total feature gain, while the 15-dimensional metadata representation contributes the remaining 21.6%. The metadata component is a much smaller part of the feature space, but its contribution is disproportionately high, indicating that metadata features are highly discriminative predictors. In particular, metadata cues tend to be the determining factors when visual artifacts are faint or subtle.

Further examination of prediction-probability distributions over datasets demonstrates that the model is context-dependent and dependent on modalities, based on the nature of underlying data.

In a dataset like CASIA v2, where the metadata is held relatively intact and informative, metadata-based predictions often provide more confidence indicators, and thus contribute significantly to the overall decision.

However, in datasets like CG-1050 and CoMoFoD, where metadata is absent, corrupted, or uninformative, the model tends to depend more on visual features, which can be seen by the dominance of visual confidence distributions.

These observations suggest that the fusion framework proposed is capable of properly leveraging the complementary strengths of both modalities. Instead of using explicit dynamic weighting, the model acquires an implicit tradeoff between visual and metadata features during training, allowing it to perform well in heterogeneous forensic settings. Both the analysis of feature importance and the visual- and metadata-modality complementary contributions to datasets are re-discovered in both the confidence distributions and the feature-importance analysis, as depicted in Figure 10.

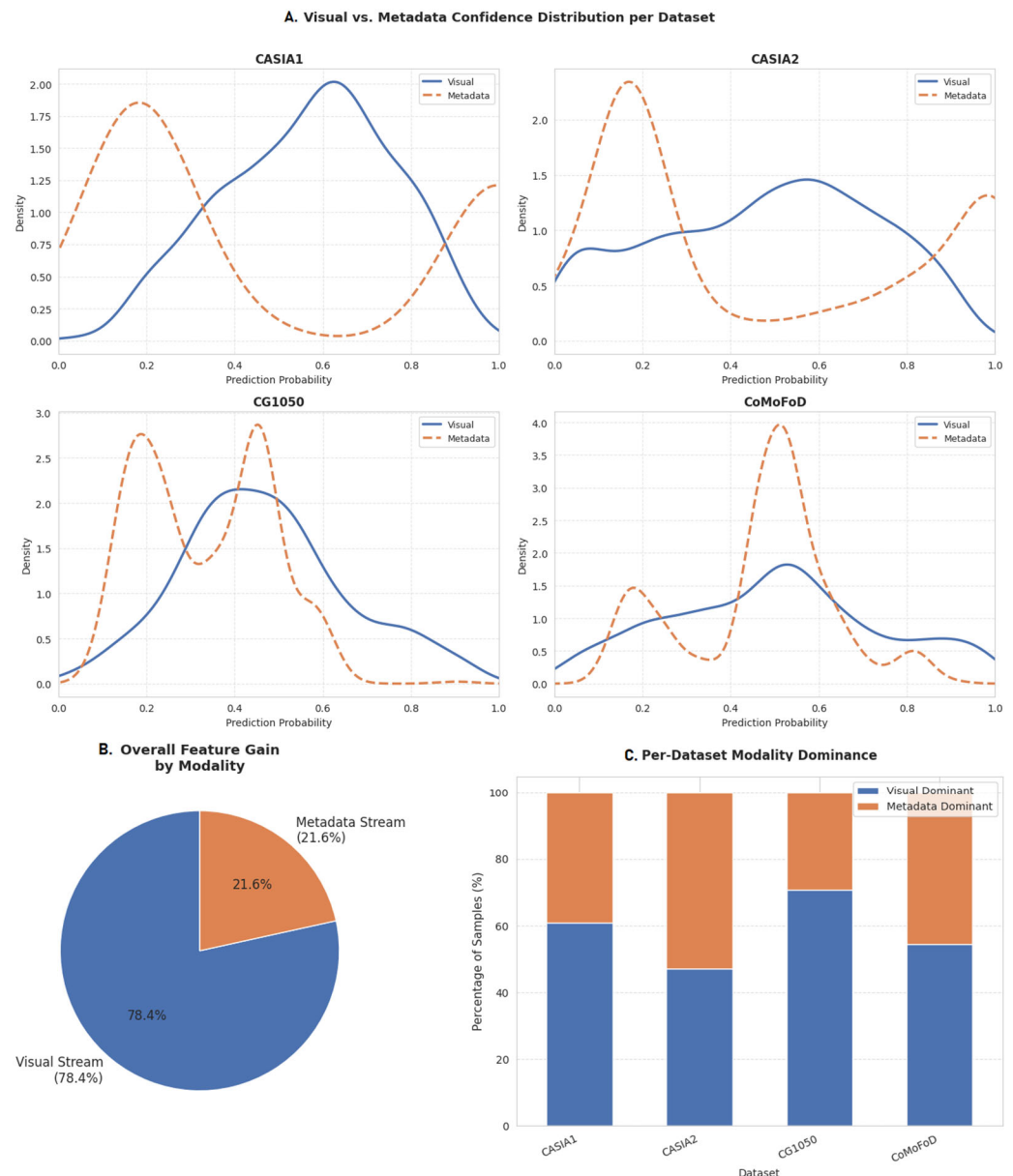


Figure 10. Confidence distributions, feature importance, and modality dominance of visual and metadata streams across datasets, demonstrating complementary and dataset-dependent behavior in the fusion model.

5.2.3. Ablation Study

Modality contribution: the ablation study measures the performance improvement of each modality on the validation split. The accuracy of the classification when only one pathway is isolated is shown in Figure 11. The baseline (MobileNetV2) achieved 68.61 accuracy using only visuals. The metadata-only baseline (XGBoost) achieved an accuracy of 75.85, which suggests the informative nature of metadata discrepancies to identify manipulation. A probability averaging of the two independent scores of probabilities raised the accuracy to 81.16. Finally, the highest accuracy of 83.85% was obtained with the proposed feature-level fusion that adds the visual embeddings to the metadata vectors, prior to the classification. This development implies that the unified classifier enjoys the advantage of learning the cross-modal associations between visual and metadata features, and not assessing them separately.

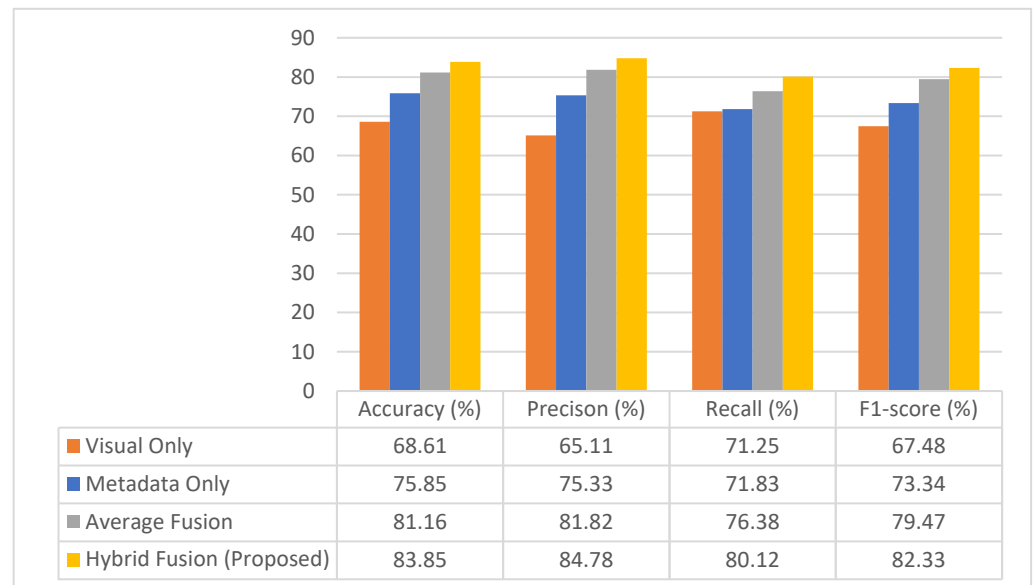


Figure 11. Ablation study detailing the contribution of independent modalities and fusion configurations.

Sensitivity to missing metadata: the study also evaluated the sensitivity of the framework to scenarios in which structural markers are completely missing, e.g., when images are processed by social media infrastructure that automatically removes EXIF tags. As a control to simulate complete data loss, we tested a zero-metadata condition by zeroing all 15 EXIF features during inference. In this state, the accuracy of the fusion model in general was reduced by approximately 30 percent. There was a 39.84% decrease in accuracy on the highly realistic CASIA v2 dataset. In contrast, datasets evaluating localized patches like CG-1050 exhibited a reduced drop of 10.59% percentage, indicating their inherently reduced dependence on camera-specific tags. This overall 30 percentage-point degradation highlights a practical limitation: while the combined metadata pathway significantly improves detection when available, the framework remains sensitive to environments where structural data is routinely removed. Figure 12 summarizes the results.

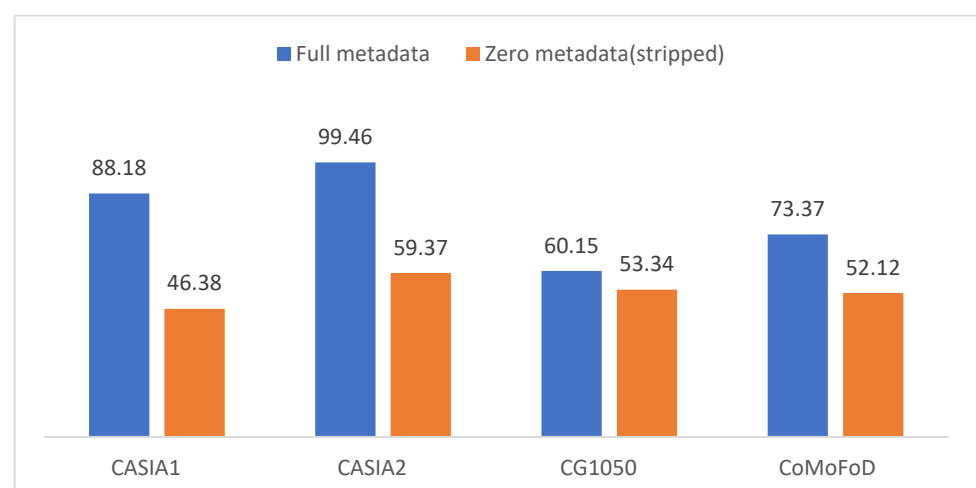


Figure 12. Full metadata vs. zero metadata (stripped) per dataset.

5.2.4. Statistical Reliability and 5-Fold Cross-Validation

After determining the diagnostic performance and component contributions of the fusion framework on our fixed test set, we further tested the model on a 5-fold cross-

validation protocol, to test absolute statistical stability. The proposed feature-level fusion model shows a high cross-fold consistency, as summarized in Table 4. The model achieved a mean accuracy of 84.17% ($\pm 0.31\%$), with a mean F1-score of 83.43% ($\pm 0.38\%$). Low inter-fold variance presents the stability of the performance across the data partitions, which supports the strength and applicability of the suggested fusion method.

Table 4. Five-fold stratified cross-validation results.

Metric	Mean (%)	Standard Deviation ($\pm\%$)
Accuracy	84.17	0.31
Precision	83.24	1.04
Recall	83.73	1.22
F1-Score	83.43	0.38

5.2.5. Cross-Modal Complementarity Analysis

To quantify the complementary contribution of the visual and metadata modalities, we conducted a best-of-both Oracle analysis on the validation split. The Oracle is a hypothetical maximum accuracy (92.43), which is defined as those cases in which at least one modality has a correct prediction.

The fusion model attains about 84.67%, reducing the gap relative to the best individual modality (75.85%). Interestingly, the fusion model manages to solve 52.8% of the cases with both modalities having uncertainty, which implies high cross-modal complementarity. In contrast, in the case of independent failure of both modalities, the fusion model only restores about 7.7% of samples, which demonstrates the poor recoverability in the absence of discriminative signals. Comprehensively, these findings indicate that the suggested framework is able to capture a significant part of the available synergy, as well as indicate potential areas of enhanced decision integration. Table 5 gives a summary of the result.

Table 5. Synergy analysis.

Metric	Value
Visual-Only Accuracy	68.61%
Metadata-Only Accuracy	75.85%
Oracle (Best-of-Both)	92.43%
Proposed Fusion	84.67%
Gap (Oracle $\rightarrow$ Fusion)	7.98 pp
Both baselines wrong	389
Fusion recovered	30 (7.7%)
Both baselines uncertain	477
Fusion correctly resolved	252 (52.8%)

5.2.6. Discussion of Implications

False information can be propagated very fast these days, particularly using manipulated images. Due to this reason, powerful forgery-detection systems are extremely significant. This paper demonstrates that it is not sufficient to conduct only visual analysis. It can happen that an image can be perfectly real, whereas its metadata (the date, time, GPS coordinates, or device data) can be altered deliberately. For example, a person may change the time of a photograph to form a fake defense or change the GPS positioning to confuse an inquiry. A system relying on visual information only would probably fail to notice this form of manipulation, but a metadata-aware system is capable of identifying such inconsistencies.

The framework, which involves a combination of visual analysis and metadata analysis, will minimize a possible risk of omitting significant evidence. It has practical significance for digital forensics, law enforcement, courts, and cybersecurity. In general, our solution assists in creating more robust digital evidence-verification frameworks at the time when manipulation of images is becoming more advanced. For real-world forensic deployment, the proposed dual-modal framework enables the efficient automated triage of large multimedia datasets. The system uses lightweight MobileNetV2 plus fast XGBoost-driven metadata analysis to label a large number of images quickly and flag suspicious images, so that they can be more thoroughly investigated by experts, which minimizes the dependence on computations and speeds up the forensic process in the workflow.

5.2.7. Ethical Considerations and Privacy

The use of automated forensic systems can have ethical issues of privacy and bias. The framework should be applied under the legal and ethical parameters, whereas it is created to identify digital manipulation.

Regarding privacy, it is important to know that metadata analysis has the ability to reveal confidential information, like information about who created the image and location information. The framework puts more emphasis on technical metadata, as compared to personally identifiable information. Unless judicial permission is given, practitioners should make sure that they adhere to the relevant privacy rules. Algorithmic bias can also exist, and the framework can be prone to biases in its training data. Trained models might be ineffective, for example, with images produced by a lower-quality device, which can potentially raise false positives. This risk was addressed by cross-dataset validation and device diversity, but continuous bias monitoring is also necessary.

5.2.8. Comparative Analysis

Table 6 gives a comparison with recent state-of-the-art methods on various datasets. The majority of the available techniques only detect visually, frequently using CNN-based architectures or improved feature representations like ELA and transfer learning. Although these methods achieve good results on certain datasets, they are usually based on one modality and are not as effective in generalizing to situations where visual artifacts are faint or hidden.

Table 6. Comparison of image-level classification accuracy with prior work on shared datasets.

Method	Modality	CASIA v1 (%)	CASIA v2 (%)	CoMoFoD (%)	CG-1050v2 (%)	Year
Narayan et al. [10]	Visual (CNN)	-	97.70	-	-	2023
Mamidwar et al. [27]	Visual (ELA + CNN Xception)	-	98.50	-	-	2024
Shirsho et al. [28]	Visual (Ensemble ELA + TL)	97.67	92.50	-	-	2024
Kaur et al. [29]	Visual (ELA + CNN + XGBoost)	90.83	96.82	-	-	2024
Bevinamarad et al. [7]	Visual (DL with ITL)	-	-	87.72	-	2025
Tallapragada et al. [30]	Visual (Enhanced Mask-RCNN)	99.70	-	-	-	2024
KASIM [31]	Visual (ResNet + PSO + GRU)	-	96.25	-	-	2024
Akshara et al. [32]	Visual (DeepLabV3)	-	99.00	-	-	2025
Proposed	Multi-modal (MobileNet-V2 + XGBoost)	88.18	99.46	73.37	60.15	2025

Note: (-) indicates that the data is not reported or not applicable.

Conversely, the suggested framework is multi-modal, combining both visual and metadata-driven forensic indications. This design is based on the expedient fact of the digital forensic investigative process, whereby pixel-based artifacts and proximal metadata are utilized to support the authenticity evaluation. By incorporating complementary sources of evidence within a unified framework, the study shifts the focus from dataset-specific optimization toward a more holistic and adaptable forensic-analysis strategy, applicable across diverse manipulation scenarios.

5.2.9. Limitations

Although the proposed framework has promising performance and implications, it is limited in a number of ways.

- ✓ The metadata module assumes that the EXIF information is intact; when metadata is intentionally washed away with anti-forensic software, the XGBoost pathway might become less effective, thereby relying on the visual features only.
- ✓ MobileNet-V2 is computationally efficient, but the framework needs large and diverse labeled datasets to be optimally fine-tuned, which may limit effectiveness in emerging scenarios involving novel generative AI techniques. The model is also not optimized specifically for the detection of AI-generated images, which may need frequency-domain or spectral analysis.
- ✓ Moreover, if both image pixels and metadata manipulation are well coordinated by an adversary, this could decrease the robustness of the model.
- ✓ Cross-dataset performance is good, but a slight decrease in accuracy on merged datasets implies that domain-shift issues still exist, with the changes in cameras and software signatures potentially influencing generalization.
- ✓ Lastly, despite a higher classification accuracy, more interpretability of internal deep-learning decisions and metadata features contributing to the final output could enhance the model's explainability.

6. Conclusions and Future Work

In this research, we have introduced a multi-modal fusion framework where the depth of convolutional neural networks is combined with the sound structured learning of the gradient-boosting machine in detecting image manipulation. The proposed MobileNet-XGBoost architecture aims to address a critical forensic need for a system that leverages all available evidence, both visual and meta-textual.

The experimental findings prove that orthogonal modalities are complementary, as visual and metadata forensics respond to various failure modes of manipulated content. The metadata-only model was more effective than the visual-only baseline, which attests to the fact that EXIF inconsistencies, including traces of editing software, are strong discriminative features. Better still, the feature-level fusion mechanism was very effective. The framework enables a high-quality input to dynamically switch between modalities due to high-level embedding concatenation, resulting in a significant performance over the single-modal approach. The fusion model showed a high recall, which means that the sensitivity of the fusion model is high when the visual traces are faint in nature. Moreover, cross-dataset generalization assessed with aggregated data sets comprising CASIA, CoMoFoD and CG-1050 shows that, despite the difficulties in heterogeneity, strong performance can still be attained under realistic forensic conditions.

Future studies will pursue two major directions. First, we will incorporate the latest Explainable AI (XAI) methods, including Grad-CAM and SHAP, to promote transparency and increase the judicial admissibility of the model's prediction. Second, we will examine how

we can use Large Language Models (LLMs) to produce automated forensic narrative reports, given any metadata anomalies identified, and thereby simplify investigative processes.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Asiri, S.; Xiao, Y.; Alzahrani, S.; Li, S.; Li, T. A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks. *IEEE Access* **2023**, *11*, 6421–6443. [\[CrossRef\]](#)
2. Li, W.; Li, B.; Zheng, K.; Li, S.; Li, H. Document image forgery detection and localization in desensitization scenarios. *Signal Process.* **2026**, *238*, 110123. [\[CrossRef\]](#)
3. Shallal, I.; Haddada, L.R.; Amara, N.E.B. Image Forgery Detection with Focus on Copy-Move: An Overview, Real World Challenges and Future Directions. *Appl. Sci.* **2025**, *15*, 11774. [\[CrossRef\]](#)
4. Chakraborty, S.; Chatterjee, K.; Dey, P. Detection of Image Tampering Using Deep Learning, Error Levels and Noise Residuals. *Neural Process. Lett.* **2024**, *56*, 112. [\[CrossRef\]](#)
5. Xu, Y.; Irfan, M.; Fang, A.; Zheng, J. Multiscale Attention Network for Detection and Localization of Image Splicing Forgery. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–15. [\[CrossRef\]](#)
6. Zheng, C.; Shrivastava, A.; Owens, A. EXIF as Language: Learning Cross-Modal Associations between Images and Camera Metadata. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 6945–6956. [\[CrossRef\]](#)
7. Bevinamarad, P.; Unki, P.H.; Bhandage, V. A deep learning model with an inductive transfer learning for forgery image detection. *Indones. J. Electr. Eng. Comput. Sci.* **2025**, *37*, 801–810. [\[CrossRef\]](#)
8. Weng, S.; Zhu, T.; Zhang, T.; Zhang, C. UCM-Net: A U-Net-Like Tampered-Region-Related Framework for Copy-Move Forgery Detection. *IEEE Trans. Multimed.* **2024**, *26*, 750–763. [\[CrossRef\]](#)
9. Yan, C.; Liu, R.; Li, H.; Wu, J.; Pan, H. Dual-Stream Intermediate Fusion Network for Image Forgery Localization. *IEEE Access* **2024**, *12*, 90511–90524. [\[CrossRef\]](#)
10. Narayan, D.; Himanshu; Kamal, R. Image Forgery Detection. In Proceedings of the 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 11–12 March 2023; pp. 549–552. [\[CrossRef\]](#)
11. Bedi, P.; Mittal, A.; Gangwar, M.; Dua, A. *Identifying Forged Images Using Image Metadata*; Springer: Cham, Switzerland, 2019. [\[CrossRef\]](#)
12. Umair, M.; Bouguettaya, A.; Lakhdari, A.; Ouzzani, M.; Liu, Y. Exif2Vec: A Framework to Ascertain Untrustworthy Crowdsourced Images Using Metadata. *ACM Trans. Web.* **2024**, *18*, 1–27. [\[CrossRef\]](#)
13. Lee, J.; Jeon, S.; Park, Y.; Chung, J.; Jeong, D. A Forensic Methodology for Detecting Image Manipulations. *arXiv* **2023**, arXiv:2308.04723. [\[CrossRef\]](#)
14. Zhong, N.; Zou, M.; Xu, Y.; Qian, Z.; Zhang, X.; Wu, B.; Ma, K. Self-Supervised AI-Generated Image Detection: A Camera Metadata Perspective. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6192–6207. [\[CrossRef\]](#)
15. Xia, X.; Su, L.C.; Wang, S.P.; Li, X.Y. DMFF-Net: Double-stream multilevel feature fusion network for image forgery localization. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107200. [\[CrossRef\]](#)
16. He, Z.; Liu, Z.; Zhao, Z. Forgery-Aware Guided Spatial-Frequency Feature Fusion for Face Image Forgery Detection. *Symmetry* **2025**, *17*, 1148. [\[CrossRef\]](#)
17. Gong, R.; Chen, J.; Zhang, D.; Sangaiah, A.K.; Alenazi, M.J.F. Face Forgery Detection via Multi-Scale and Multi-Domain Features Fusion. *IET Image Process.* **2025**, *19*, e70131. [\[CrossRef\]](#)

18. Umair, M.; Singh, P.; Bouguettaya, A. A Context-Aware Service Framework for Detecting Fake Images. In *Service-Oriented Computing*; Gaaloul, W., Sheng, M., Yu, Q., Yangui, S., Eds.; Springer Nature: Singapore, 2025; pp. 296–311. [\[CrossRef\]](#)
19. Liu, C.; Yao, W. MMFA-Net: Multi-Modal Feature Learning and Multi-Head Fuse Attention for Image Manipulation Detection and Localization. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*; IEEE: Shenzhen, China, 2024; pp. 1298–1302. [\[CrossRef\]](#)
20. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [\[CrossRef\]](#)
21. Shang, J.-J.; Phipps, N.; Wey, I.-C.; Teo, T.H. A-DSCNN: Depthwise Separable Convolutional Neural Network Inference Chip Design Using an Approximate Multiplier. *Chips* **2023**, *2*, 159–172. [\[CrossRef\]](#)
22. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [\[CrossRef\]](#)
23. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
24. Dong, J.; Wang, W.; Tan, T. CASIA Image Tampering Detection Evaluation Database. In *Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing*, Beijing, China, 6–10 July 2013; pp. 422–426. [\[CrossRef\]](#)
25. Tralic, D.; Zupancic, I.; Grgic, S.; Grgic, M. CoMoFoD—New database for copy-move forgery detection. In *Proceedings of the ELMAR-2013*, Zadar, Croatia, 25–27 September 2013; pp. 49–54. Available online: <https://ieeexplore.ieee.org/document/6658316> (accessed on 13 February 2026).
26. Castro, M.; Ballesteros L, D.M.; Renza, D. *CG-1050 v2: Original and Tampered Images, Version 1*; Universidad Militar Nueva Granada: Sede Bogotá, Colombia, 2019. [\[CrossRef\]](#)
27. Mamidwar, A.; Mugalikar, A.; Ghuge, V.; Godse, D. Image Forensics—Analysis of Manipulated Images. In *Innovations in Computational Intelligence and Computer Vision*; Roy, S., Sinwar, D., Dey, N., Perumal, T., R.S. Tavares, J.M., Eds.; Springer Nature: Singapore, 2024; pp. 371–391. [\[CrossRef\]](#)
28. Shirsho, M.H.; Rana, M.M.; Akhter, J.; Rahaman, A.S.M.M. Enhancing Image Manipulation Detection through Ensemble ELA and Transfer Learning Techniques. *Int. J. Comput. Digit. Syst.* **2025**, *18*, 1–12. [\[CrossRef\]](#)
29. Kaur, S.; Sinha, N.; Jain, P.; Koli, S.; Sharma, A.; Lathwal, A. Enhanced Image Forgery Detection using a Hybrid Approach: Integration of ELA, CNN, and XGBoost. *Int. J. Perform. Eng.* **2024**, *20*, 367. [\[CrossRef\]](#)
30. Tallapragada, V.V.S.; Reddy, D.V.; Kumar, G.V.P. Blind forgery detection using enhanced mask-region convolutional neural network. *Multimed. Tools Appl.* **2024**, *83*, 84975–84998. [\[CrossRef\]](#)
31. Kasim, Ö. Deep learning-based efficient and robust image forgery detection. *Multimed. Tools Appl.* **2024**, *83*, 59819–59838. [\[CrossRef\]](#)
32. Akshara, A.; Gayathri, R.; Alex Raj, S.M. Leveraging DeepLabV3 for Accurate Localization of Image Forgeries. In *Proceedings of the 2025 6th International Conference on Control, Communication and Computing (ICCC)*, Thiruvananthapuram, India, 23–25 May 2025; pp. 1–5. [\[CrossRef\]](#)

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.