mathematics

MDPI

Article

One Model for Many Fakes: Detecting GAN and Diffusion-Generated Forgeries in Faces, Invoices, and Medical Heterogeneous Data

Mohammed A. Mahdi [1], Muhammad Asad Arshed [2,*] and Amgad Muneer [3]

[1] Information and Computer Science Department, College of Computer Science and Engineering, University of Ha'il, Ha'il 55476, Saudi Arabia; m.mahdi@uoh.edu.sa
[2] School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan
[3] Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Perak, Malaysia; amgad_20001929@utp.edu.my
* Correspondence: asad.arshed@umt.edu.pk

Abstract

The rapid advancement of generative models, such as GAN and diffusion architectures, has enabled the creation of highly realistic forged images, raising critical challenges in key domains. Detecting such forgeries is essential to prevent potential misuse in sensitive areas, including healthcare, financial documentation, and identity verification. This study addresses the problem by deploying a vision transformer (ViT)-based multiclass classification framework to identify image forgeries across three distinct domains: invoices, human faces, and medical images. The dataset comprises both authentic and AI-generated samples, creating a total of six classification categories. To ensure uniform feature representation across heterogeneous data and to effectively utilize pretrained weights, all images were resized to 224 × 224 pixels and converted to three channels. Model training was conducted using stratified K-fold cross-validation to maintain balanced class distribution in each fold. Experimental results of this study demonstrate consistently high performance across three folds, with an average training accuracy of 0.9983 (99.83%), validation accuracy of 0.9620 (96.20%), and test accuracy of 0.9608 (96.08%), along with a weighted F1 score of 0.9608 and exceeding 0.96 (96%) for all classes. These findings highlight the effectiveness of ViT architectures for cross-domain forgery detection and emphasize the importance of preprocessing standardization when working with mixed datasets.

Keywords: cross-domain forgery; GAN and diffusion models; multiclass; vision transformer; pretrained models; stratified K-fold; deep learning

MSC: 68T07

Academic Editor: Haifeng Wang

Received: 27 August 2025
Revised: 20 September 2025
Accepted: 22 September 2025
Published: 26 September 2025

Citation: Mahdi, M.A.; Arshed, M.A.; Muneer, A. One Model for Many Fakes: Detecting GAN and Diffusion-Generated Forgeries in Faces, Invoices, and Medical Heterogeneous Data. *Mathematics* **2025**, *13*, 3093. https://doi.org/10.3390/math13193093

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

1. Introduction

The current advancement of Artificial Intelligence (AI) technologies has transformed several sectors, including image processing, document processing, and finance [1]. While these innovations have increased operational efficiency, they have also caused new vulnerabilities, especially in financial fraud, including events like accounting fraud and insider trading [2]. Nowadays, one emerging threat is the creation of AI-generated media, which can be used to falsify records and facilitate fraud. These AI-generated media closely resemble real media, making it increasingly difficult to distinguish between them due to advanced image generation techniques [3].

In the medical domain, CT scans, X-ray Scans, MRIs, and ultrasound play an active role in disease diagnosis. According to a Harvard Health report [4], approximately 80 million scans are performed annually in the United States. In the past, image tampering mainly targeted human faces and video, but today, other domains such as medical imaging [5] and finance [6] are also being affected by deepfakes.

Invoice fraud has become a serious concern, moving from a theoretical concept to a frequent real-world threat, especially in accounts payable operations. According to a 2024 survey, finance teams in the USA and the UK experience 13 invoice frauds per month on average, which cause an average loss of $133,000 and £104,000 in the USA and the UK, respectively [7]. The severity of this issue is highlighted by numerous high-profile cases. For instance, the National Trust lost £1 million due to an internal scam perpetrated by an employee who authorized 148 fake invoices submitted by his son [8]. Similarly, an employee in Poland generated falsified VAT invoices to facilitate fraudulent tax claims [9], and a person in Fort Lauderdale fell victim to a $1.2 million phishing scam involving a fraudster mimicking a local contractor [10]. These examples underscore the major internal and external threats posed by advanced AI forgery techniques, making detection increasingly complex. There are several image forgery techniques available, the most common of which are copy-move, Splicing, Inpainting, DeepFakes, Computer-Generated Imagery (CGI), and Generative Adversarial Network (GAN) -based face synthesis [11].

1.1. AI Forgery Methods to Improve Readability

Copy-Move Method: In the Copy-move forgery method, to duplicate or cover the elements, a part of the image is copied and pasted onto another area of the image. This technique is commonly used for tampering purposes due to the similar properties of the source and the destination. In 2024, Shinde et al. [12] proposed an efficient method for copy-move forgery detection using a Graph Convolutional Network (GCN) with multiple layers and a well-known ReLU activation function. They achieved a 99% validation accuracy for the MICC F220 Dataset [13] with their proposed detection method.

Splicing: In the splicing forgery method, the components of two or more images are combined into a composite image to manipulate the visual content. The detection of splicing forgery is more challenging than the copy-move technique due to external elements. In 2024, Yang et al. [14] proposed a dual encoder network model named D-Net for splicing forgery detection. Their study was based on forensic fingerprints, as compared to the previous detection method that relied on semantic features.

Inpainting: In this forgery technique, digital images are manipulated by removing specific image regions and filling them with surrounding pixels. Li et al. [15] proposed an approach for inpainting forgery detection using label decoupling and constrained adversarial training. The detection performance increased in their study by the introduction of inference noise in the inpainted regions.

DeepFakes: Deepfakes represent media that look real, but these are generated with artificial intelligence (AI) algorithms. Identifying such media from the human eye is challenging due to the rapid advancement of AI image generation technologies. Deepfakes are based on neural networks and require massive datasets to learn face features, voice, and expressions, enabling the creation of images or videos. Nowadays, anyone can create fake media using openly available tools and models like FaceDancer [16], OpenAI's Sora [17], and Google's Veo 2 [18]. The process of creating fake images is heavily based on an autoencoder that is trained on real and altered images/videos. Initially, the encoder learns the differences between real and deepfake and, for each type, produces an equivalent latent representation. Similarly, the decoder part used the latent to reconstruct the prelimi-

nary input. The process of generating such effective deepfake content relies on different technologies, including 3D (ResNet and ResNeXt) [19].

Goodfellow et al. [20] developed the first GAN model. The GAN is rapidly being used to create deepfake content, such as fake images [21], and GANs are considered unsupervised as generative models due to their ability to locate the distribution of data automatically. The two primary components of the GAN are the Generator and Discriminator, which are used to generate data and classification, respectively, see Figure 1.

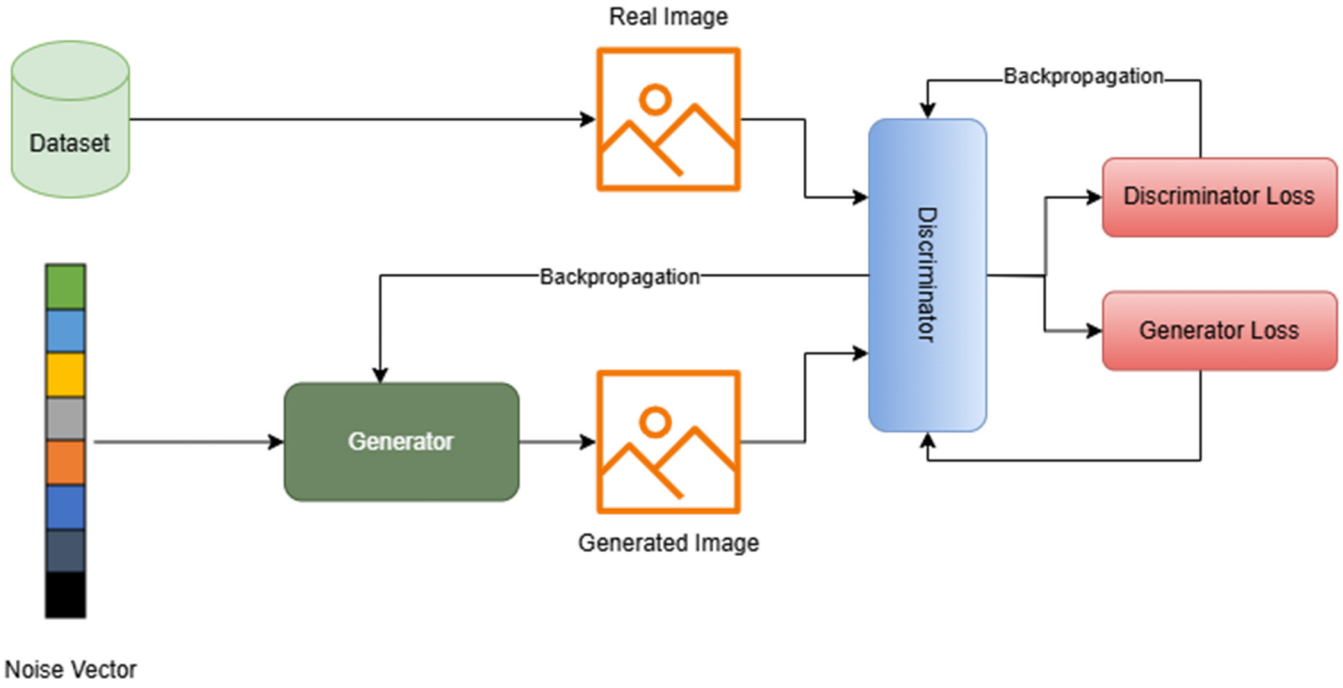


Figure 1. Abstract Architecture of GAN.

The GAN employs a supervised learning concept for synthetic data generation, which enables it to engage in unsupervised learning. The GAN model assumes structures are available in the dataset without considering findings (labeled or classified). The GAN formula is available in Equation (1) [20].

$$min_G max_D V\ (G, D) = \left[E_{x \sim P_{data}(x)}[logD(x)] + E_{z \sim P_z(z)}[log(1 - D(G(z)))]\right] \quad (1)$$

In Equation (1), G, D, $P_{data}(x)$, $P_z(z)$, $D(x)$, $G(z)$ denote the generator, discriminator, real data distribution, prior noise distribution, a discriminator output, and generator output, respectively. Similarly to the GAN, the stable diffusion model, released in 2022 [22], is also being considered for deepfake generation, particularly in the context of text prompts. This model was initially designed for artistic and creative purposes. With these advanced features, it is easy to manipulate images and video to present false information. The generation of an image from a text prompt includes several steps, like text encoding, synthesis, and refinement. The text encoding transforms the text into a format that the next synthesis model can easily process. The process of Stable Diffusion text-to-image conversion typically encompasses several key steps, including text encoding, image synthesis, and refinement. During text encoding, the textual descriptions transform into a format compatible with processing by the image synthesis model. Text embeddings or attention mechanisms capture the semantic meaning. Stable diffusion is based on the Latent Diffusion Model (LDM), where the image is generated from text conditions in latent space rather than pixel space.

This feature can help to minimize the computation cost without compromising the image quality. The random noise is refined iteratively into the image with the reverse diffusion process [22], the LDM core loss equation can be seen in Equation (2). Each step of LDM learns a shared denoising autoencoder function denoted as $\in_{\vartheta} (z_t, t, T_{\vartheta}(y))$, conditioned on the timestep t and is shared across all diffusion steps $t = 1 \ldots T$.

$$\text{LDM} = E_{z_t,\, y,\, \epsilon \sim N(0,I),t} \left[||\in - \in_{\vartheta}(z_t, t, T_{\vartheta}(y))||_2^2 \right] \quad (2)$$

1.2. Research Gap and Contributions

Although numerous studies exist for deepfake detection across various domains, including medical and social media, a notable gap remains in research related to the finance domain, as well as a single model for cross-domain detection. Due to advancements in text-to-image generation models [23], there is also a high risk of generating realistic yet fake financial documents, as well as medical scans and human faces. The latest models have not been thoroughly explored, and no single model of identification has been proposed yet for fake invoices, medical and faces. In this study, we utilized the newest version of ChatGPT-4 [24] to generate a fake invoices dataset. For other domain datasets (including real invoices, medical and human faces), the public repository is considered in this study. Furthermore, we propose a single deep learning model capable of distinguishing between fake and real invoices, human faces, and medical images. The main contributions of this paper are summarized as follows:

- We propose a single, unified deep learning model capable of identifying forgeries across three highly diverse domains—financial invoices, medical CT scans, and human faces—eliminating the need for domain-specific models.
- We introduce a new composite dataset for cross-domain forgery detection. This dataset combines publicly available images with a novel set of realistic, fake invoices generated using a state-of-the-art text-to-image model (GPT-4), providing a valuable resource for future research.
- We demonstrate that a patch-based approach can learn generalizable, domain-invariant features, proving effective for detecting manipulations in structurally distinct image types (documents, medical scans, and natural images).
- We provide an extensive comparative analysis, showing that our proposed model outperforms standard CNN-based pretrained models, establishing a new benchmark for cross-domain forgery detection.

The remaining parts of the study are structured such that Section 2 provides an overview of the literature, covering different areas where deepfakes are being used, as well as the detectors that have recently been introduced. Sections 3–5 of this study present the methodology, experiments, and conclusion.

2. Literature Review

The rapid advancement of computer vision, especially with models like CNNs and transformers, is significant. Consequently, the use of deepfakes to create AI-generated images has become increasingly common. These technologies are also being used for developing advanced medical deep learning models. GANs are being used to artificially expand datasets, overcoming the limitations of medical datasets [25]. In the study [25], the authors observed a significant performance improvement (sensitivity increased from 78.6% to 85.7%) in a CNN model using GAN-based synthetic images. While these techniques offer benefits such as dataset augmentation, they also enable malicious activities like image tampering. New detectors for these forgeries are also being proposed regularly [26].

In the medical domain, Amiri et al. [27] proposed a model for detecting copy-move forgery. They considered a dataset of 300 images, comprising 200 fake and 100 authentic images [28], and achieved an accuracy of 90.07%. A MedNet model was proposed by Albahli et al. [29] for the detection of CT scan lung fake images. Their model, based on the EfficientNetV2-B4 [30] pre-trained model, achieved an accuracy of 85.49%.

In 2019, Akhtar and Dasgupta [31] proposed that manipulated faces can be identified using the local-feature descriptors. In their study, they considered the dataset [32] named 'deepfakeTIMIT' and ten different descriptors.

In 2022, Sharafudeen et al. [32] considered a 3-dimensional neural network for the identification of tampered and original CT scan lung images. Another study by Budhiraja et al. [33] highlighted the primary threat of deepfakes, particularly in the medical domain, where tampered images can be used to mislead diagnoses. They proposed reservoir computing combined with a convolutional feature extraction technique to effectively capture temporal information in medical images, particularly CT scans.

In 2024, Sharafudeen et al. [34] utilized the CGAN [35] to generate synthetic skin lesion images and employed the Vision Transformer (ViT) model for the detection of these images. They achieved an accuracy of 97.18% in their study. Zhang et al. [36] proposed a two-stage cascade framework to identify small-region forgeries, especially those generated by GANs, such as CT-GAN. Their approach is based on local detection and can be particularly helpful for detecting subtle manipulations. Their proposed method addresses the challenge of identifying the minute tampering, which is based on less than 1% of the original image. Arshed et al. [37] proposed a fine-tuned pre-trained model to differentiate between real and fake malignant images. They prepared the fake malignant images using stable diffusion [22]. They achieved accuracy, precision, recall, and F1 of 99.66%. They also considered transformer-based architecture in their study, which achieved an accuracy, precision, recall, and F1 score of 99.66%. They also conducted human tests to demonstrate the robustness and necessity of the AI model.

Bekci et al. [38] proposed a detection system based on steganalysis models and metric learning to effectively identify different manipulations on unseen samples effectively. In their study, they considered three well-known public datasets: Celeb DF [39], FaceForensics++ [40], and DeepfakeTIMIT [41], and achieved accuracy improvements of 5–15% for unseen manipulation.

The eyebrow region was identified as a key feature in the study by Nguyen et al. [42]. They applied different CNN models such as LightCNN, ResNet, DenseNet, and SqueezeNet. They achieved an AUC score of 0.984 for the dataset UADFV [43] and 0.712 for the dataset Celeb-DF [39].

The algorithms, especially forensic algorithms, are mainly based on human input and DL for the identification of content authenticity [44]. Silva et al. [44] considered attention-based methods for deepfake identification, demonstrating how these methods identify faces and other image components.

3. Materials and Methods

This section provides a detailed description of the datasets and methodologies used to identify fake cross-domain media. The overall methodology, illustrated in Figure 2, consists of four main stages: (1) preparation of a cross-domain dataset from faces, invoices, and medical scans, (2) data preprocessing, (3) a Transformer-based model for classification, and (4) model training and evaluation using 3-fold stratified cross-validation. These methods are applied to three critical domains: preventing financial fraud by detecting fake invoices and counterfeit refund claims; ensuring medical integrity by identifying fake medical scans

to prevent fraudulent insurance claims; and verifying user authentication through facial recognition.

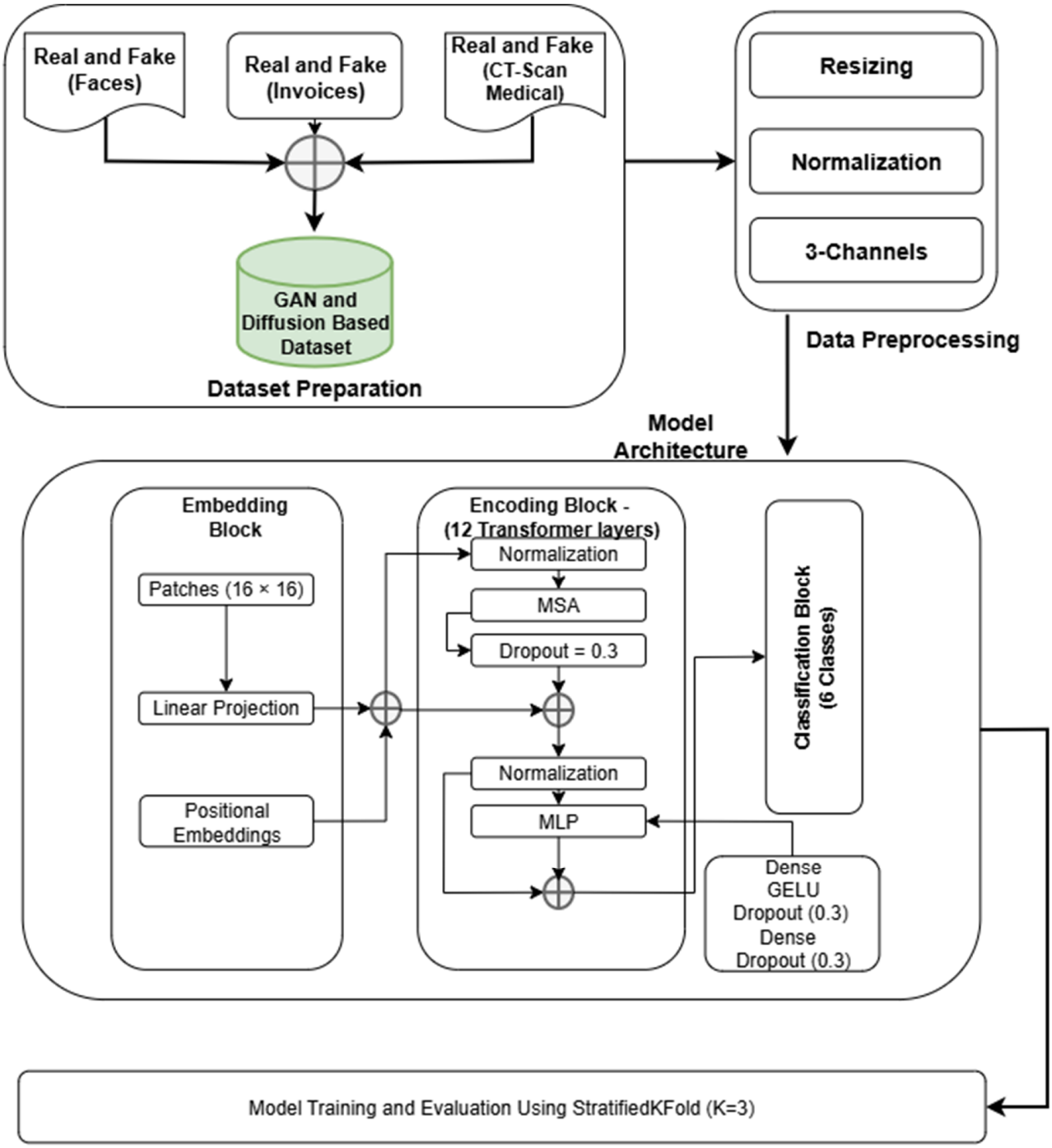
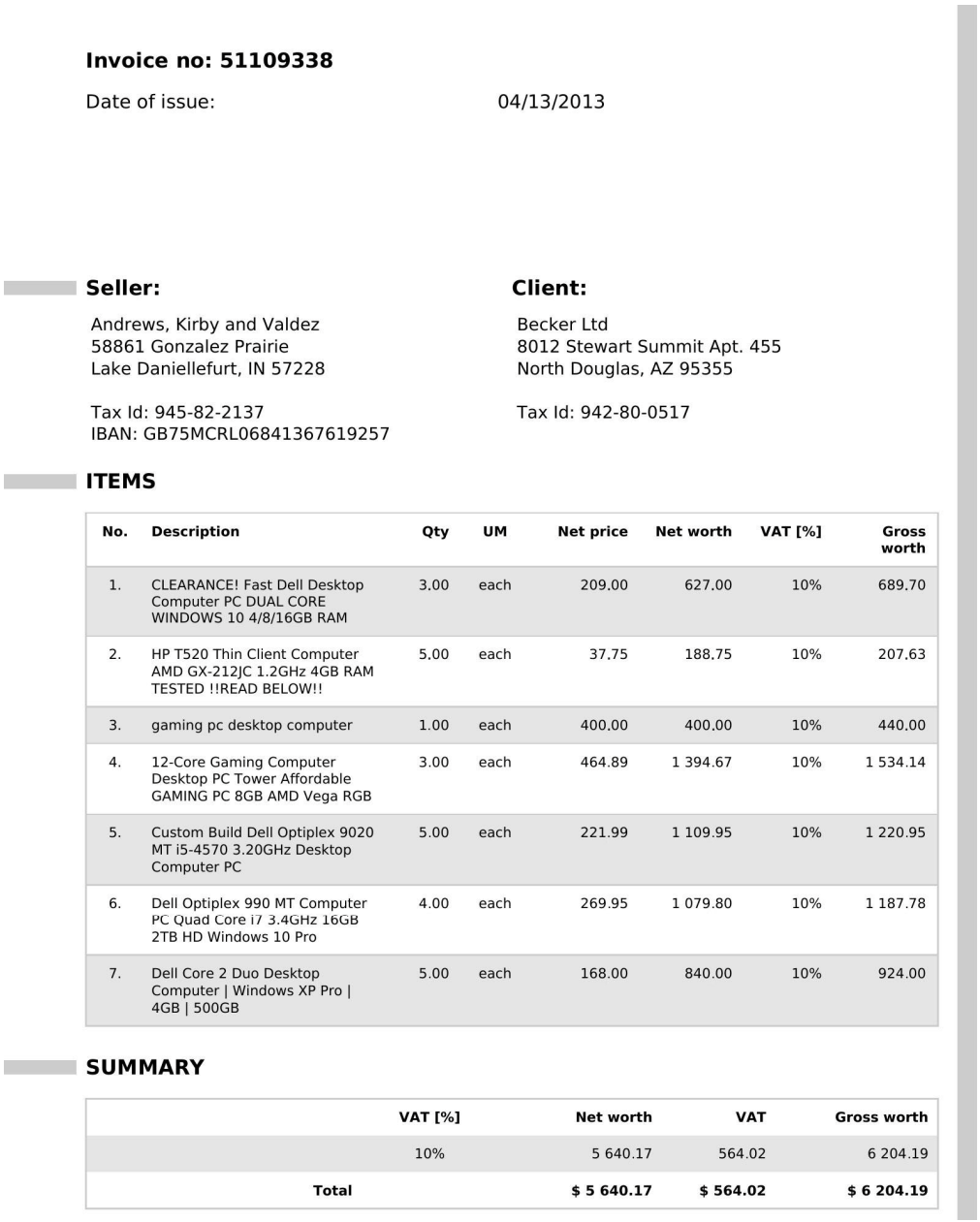


Figure 2. The overall architecture of the proposed system for cross-domain forgery detection.

3.1. Dataset Description

To develop a robust deep learning model for identifying fake invoices, a custom dataset was assembled for this study, comprising two classes: Authentic and Fake (ChatGPT). To reflect the diversity of real-world invoices, various factors were considered during dataset creation, including layout variations and visual presentation.

Real and Fake Invoices: In this study, for the real invoices, the open-source dataset named "High-Quality Invoice Images for OCR" available at Kaggle is considered [45]. This dataset consists of high-quality images; sample images from the dataset are shown in Figure 3. We have considered the 265 images from this dataset due to computational limitations.



Invoice no: 51109338

Date of issue: 04/13/2013

Seller:

Andrews, Kirby and Valdez
58861 Gonzalez Prairie
Lake Daniellefurt, IN 57228

Tax Id: 945-82-2137
IBAN: GB75MCRL06841367619257

Client:

Becker Ltd
8012 Stewart Summit Apt. 455
North Douglas, AZ 95355

Tax Id: 942-80-0517

ITEMS

No.	Description	Qty	UM	Net price	Net worth	VAT [%]	Gross worth
1.	CLEARANCE! Fast Dell Desktop Computer PC DUAL CORE WINDOWS 10 4/8/16GB RAM	3.00	each	209.00	627.00	10%	689.70
2.	HP T520 Thin Client Computer AMD GX-212JC 1.2GHz 4GB RAM TESTED !!READ BELOW!!	5.00	each	37.75	188.75	10%	207.63
3.	gaming pc desktop computer	1.00	each	400.00	400.00	10%	440.00
4.	12-Core Gaming Computer Desktop PC Tower Affordable GAMING PC 8GB AMD Vega RGB	3.00	each	464.89	1 394.67	10%	1 534.14
5.	Custom Build Dell Optiplex 9020 MT i5-4570 3.20GHz Desktop Computer PC	5.00	each	221.99	1 109.95	10%	1 220.95
6.	Dell Optiplex 990 MT Computer PC Quad Core i7 3.4GHz 16GB 2TB HD Windows 10 Pro	4.00	each	269.95	1 079.80	10%	1 187.78
7.	Dell Core 2 Duo Desktop Computer | Windows XP Pro | 4GB | 500GB	5.00	each	168.00	840.00	10%	924.00

SUMMARY

	VAT [%]	Net worth	VAT	Gross worth
	10%	5 640.17	564.02	6 204.19
Total		**$ 5 640.17**	**$ 564.02**	**$ 6 204.19**

Figure 3. Sample of Original Invoice [45].

The fake receipt was generated using the integrated image generation ability of DALL-E 3 [46] (proprietary text-to-image diffusion model) by OpenAI [24]. With careful prompt engineering, the model generated a high-resolution image that resembles a realistic receipt. The generated image samples are shown in Figure 4, and we have created a total of 265 images.

Invoice nu: 91111647

Date of issue: 12/01/2014

Seller:

Gamex-Electronics Ltd.
90 Pretty Pine Ct.
Belieview, OH 44809

Tax Id: 352-73-7591

Client:

Highpoint Ltd
814 Vailey View Farge Apt. 097
Horshoim, MS 803S4

Tax Id: 067-83-8604

ITEMS

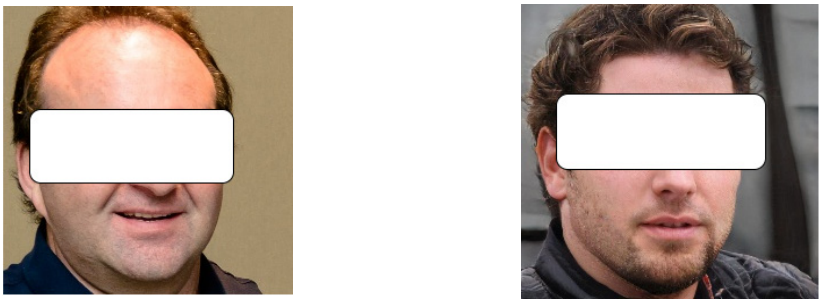
No.	Description	Qty	UM	Net price	Net worth	VAT [%]	Gross worth
1.	CLEARANCI¶ Fast Deill Dektop Complute PC WIN To PENTIUN DoAL CORE	4	each	$229.00	$91.80	10%	$100.00
2.	HP T530 Thin Client Computer AND ES 4:.)2 of)56 PAM TESTED ⁊HRAEDEER	4	each	$460.00	$80.00	10%	$80.00
3.	gaming pc desktop computer	2	each	$644.99	$1.94.87	10%	$139.41
4.	12 Core Gaming Computer Desktop PC Tower Afforciable GAMING PC Tower	3	each	$221.99	$10.03	10%	$120.03
3.	Custom Build Deil Optiplex 9020 M-1-9-4572 3 30GHz Desktop Computer PC	5	each	$221.89	$1,03.35	10%	$107.69
6.	Dell Optiplee 939 MT Computer PC Ouad Core X 5-47GHz 15GB 2TD HD Windows 3 Pro	4	each	$280.85	1.079.00	10%	$197.79
7.	Deil Core 2 Dud Desktop Computer 1 Windows XF Pro | 4GB | 500GE	5	each	$840.00	$84.00	10%	$624.00

SUMMARY

VAT [%]	Net worth	VAT	Gross worth
10%	$8.491.42	$43.14	$349.14
Total	**$ 5 491.42**	**$ 549.14**	**$ 6 040.56**

Figure 4. Sample of Fake Invoice.

Real and Fake Faces: The real and fake faces images were retrieved from the open-source dataset available at Kaggle [47]. The real images of this dataset are based on the Flickr dataset collected by Nvidia, while the fake images were generated by StyleGAN [48]. Sample images from this dataset are shown in Figure 5.



Real **Fake**

Figure 5. Sample images of authentic and fake faces from the dataset [47].

Real and Fake CT-Medical Images: The authentic and fake CT scan images were sourced from an open-source dataset [49]. This dataset provides authentic CT scans alongside two categories of fake images generated by a diffusion model: those with maliciously injected artifacts and those where original features have been removed. Sample images illustrating these categories are shown in Figure 6.

Figure 6. Sample of Real and Stable Diffusion-Based Injected Fake Image [49].

The final dataset is a comprehensive, multi-domain collection specifically assembled to train and evaluate a robust forgery detection model. It integrates authentic images from public repositories with a diverse set of forgeries generated using state-of-the-art models, including StyleGAN [48], diffusion models [46], and GPT-4 [24]. This results in a final dataset of six distinct classes (Authentic and Fake for each of the three domains: invoices, faces, and medical scans). Each class contains 265 images, with the exception of the Real-Medical class, which has 207 images.

3.2. Dataset Preprocessing

The images of the dataset are based on a high-quality resolution, even typically for some classes exceeding (1024, 1024). While such resolutions offer rich visualization, processing them directly is a significant challenge due to memory and computation time constraints, especially when working with limited resources. To enable efficient training of the deep learning models, such as the DenseNet201 [50] pretrained model, the following preprocessing steps were taken in this study.

- Resolution Normalization: The complete dataset of this study was resized to a 224×224 fixed resolutions. The images were resized to this size due to the standard size of available Convolutional Neural Network (CNN) based pretrained models, such as MobileNetV2 [51] and VGG16 [52]. These resizing steps help reduce memory usage slightly during training and inference.
- Pixel Value Scaling: The pixel values were normalized with a range [0, 1] for the stable gradient updates and faster convergence during the model training.
- Removed Images: We have also removed images that ChatGPT does not properly generate (some text-like items' descriptions are misprinted).
- Conversion to 3-Channels: To make the dataset consistent and apply pretrained weights, the dataset was converted to 3-Channels.

3.3. Dataset Average Intensity Histogram Analysis

When the dataset consists of mixed RGB and Grayscale images, keeping them in their original format can lead to inconsistency in feature representation, and this can negatively affect the model's performance. Most deep learning models, such as ViT, require uniform input dimensions and channels. Suppose grayscale images are left as single channels and RGB

images as three channels. In that case, the model will learn different statistical distributions for the same feature space, which can confuse the early layers and reduce generalization.

Although the fake images appear realistic to the human eye, pixel-level intensity differences can be observed during the analysis phase [53]. To explore this, we computed the average intensity histograms for real and fake classes after converting the dataset into RGB to ensure consistency and eliminate color channel bias. As shown in Figure 7, the average pixel intensities for authentic and fake invoices are highly similar. In the medical domain, however, a significant discrepancy is visible: authentic medical images have a much higher average intensity, indicating that the fake images are generally darker. A similar but less pronounced difference is observed in the face domain, where the authentic and fake images exhibit closely aligned intensity distributions with only minor variations.

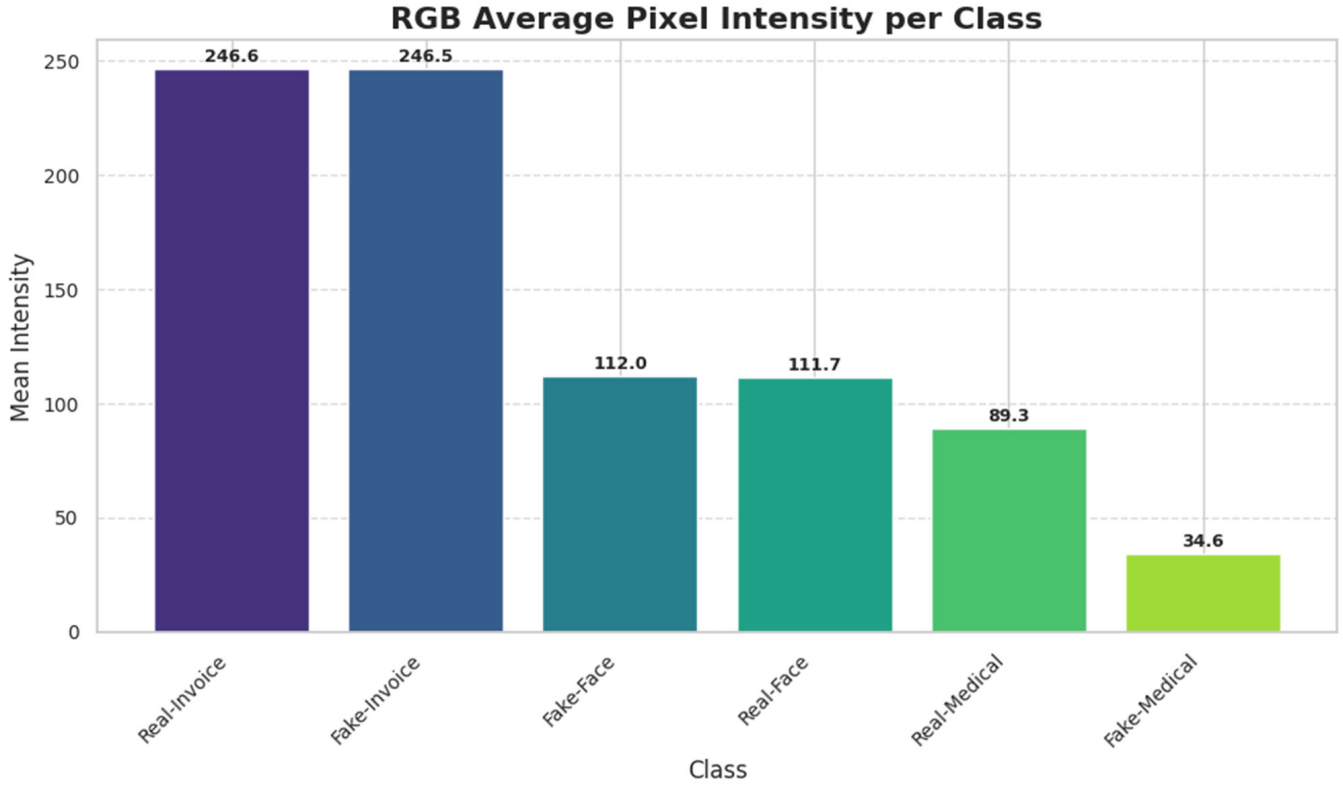


Figure 7. Comparison of average pixel intensity histograms for authentic and fake images across the three domains (invoices, medical, and faces) after RGB conversion.

3.4. PCA-Based Latent Space Analysis

To investigate the separability of real and fake images across domains, we performed Principal Component Analysis (PCA) [54] on RGB versions of all samples to ensure uniformity. Each image was resized to (224, 224) and flattened into a 1D vector of 50,176 features before projection into a 2D latent space.

As shown in Figure 8, distinct clusters emerge corresponding to the six classes: Real-Face, Fake-Face, Real-Medical, Fake-Medical, Real-Invoice, and Fake-Invoice. The Fake-Invoice class forms a highly compact cluster, suggesting low intra-class variation—possibly due to consistent patterns in AI-generated content. In contrast, courses like Real-Face and Fake-Face display wider spatial spread, reflecting natural variability and generation diversity.

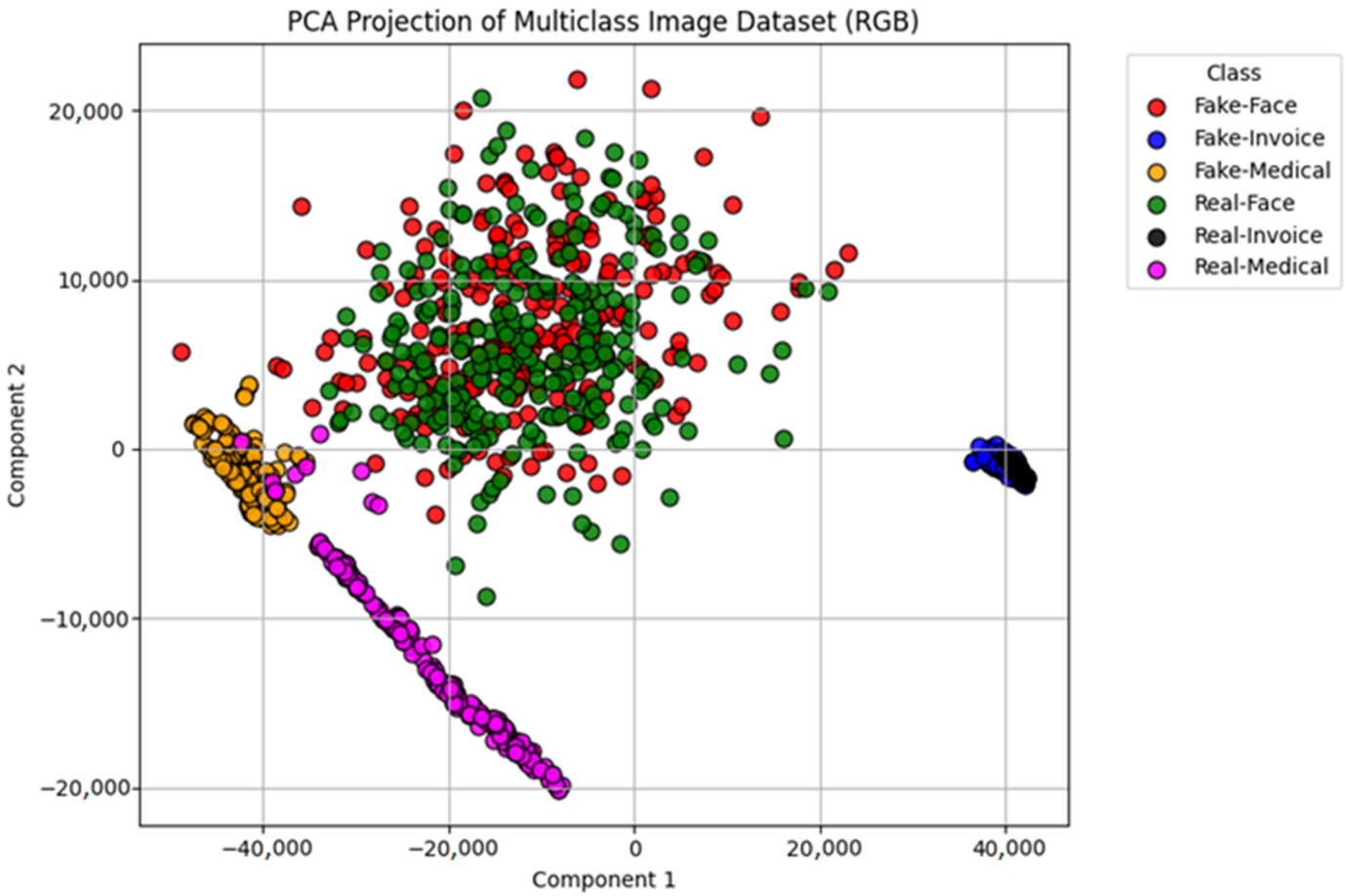


Figure 8. PCA Projection of the Multiclass Image Dataset across Six Categories (RGB).

This PCA projection confirms that even low-dimensional statistical representations carry meaningful class-wise separability. It reinforces the idea that real and fake examples—despite looking visually similar—occupy distinct regions in feature space. These findings provide valuable cues for designing automated classifiers based on latent embeddings.

3.5. Proposed Model and Hyperparameters

The ViT model, introduced in 2020 [55], is best suited for image analysis tasks in computer vision. Similarly to the use of transformers in NLP tasks, ViT also treats images as a sequence of tokens and commonly represents them as patches. The essential parts of ViT architecture are embedding and tokenization. It divides the input image into a grid of non-overlapping patches. With linear operations, these patches are transformed into higher-dimensional space. This enables the ViT model to extract both local and global information from the image [56].

The ViT model architecture is primarily composed of four components: patch embedding, positional embedding, transformer encoder, and classification head. In the embedding patch, the image patches are linearly projected to the embedding space with a linear transformation. In the next step, to capture the spatial relationship, each patch embedding is augmented with positional information. The transformer encoder processes the positional embeddings. The encoder comprises several layers featuring a self-attention mechanism and a feed-forward network. The overall result of this encoder component is a contextualized embedding. The final classification head block takes contextualized embedding and, with one or more fully connected layers, it makes the final prediction. Furthermore, to capture the diverse patterns, the multi-head self-attention (MSA) enables the ViT model to

focus on the different parts of the image simultaneously. The scaled dot product attention is calculated using Equation (3), in which Q, K, V = Queries, Keys, and Values matrices, and d_k represents the dimension of the keys.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

In this study, we observed a mixed dataset, with some classes in grayscale and others in RGB. To ensure consistency, we converted all images to RGB. Furthermore, all the images are resized to 224 × 224, and the patch size is 16 × 16. In the patch embedding, there will be 196 (14 × 14) patches, calculated as (224/16 × 224/16), each with dimensions 16 × 16 × 3.

Furthermore, each 16 × 16 pixel patch is flattened into a 1D vector, which is then passed through a linear layer to generate a fixed-size patch embedding of 768 dimensions. A learnable Classification (CLS) token is then prepended to the sequence of patch embeddings; this special token serves to aggregate a global representation of the entire image. To retain spatial information, which transformers inherently lack, positional embeddings are added to each embedding, resulting in a total of 197 embeddings (196 from patches + 1 CLS token). This sequence is passed through 12 encoder layers, each with 12 attention heads, to capture rich contextual relationships. Each encoder layer includes a Multi-Head Self-Attention (MSA) block, a feed-forward network, residual connections, layer normalization, and a dropout rate of 0.3. Finally, the output embedding corresponding to the CLS token is passed through a fully connected linear layer to produce the final prediction across the six classes. The specific hyperparameters are detailed in Table 1.

Table 1. Proposed Model Hyperparameter Configuration.

Hyperparameter	Value
Input Image Size	224 × 224
Patch Size	16 × 16
Number of Patches	196
Embedding Dimension	768
[CLS] Token	Added (Prepended)
Positional Encoding	Learnable (197 × 768)
Number of Transformer Layers	12
Number of Attention Heads	12
FFN Hidden Size	768 × 6 = 4608
Dropout Rate	0.3

4. Results and Discussion

This section presents the experimental results of our proposed model and benchmarks its performance against several baseline models using standard evaluation metrics.

4.1. Evaluation Metrics

Evaluation metrics, such as accuracy, precision, recall, and F1 score, play a crucial role in DL model evaluation. The performance of the AI models was checked using evaluation metrics.

- Accuracy: The overall model's correctness prediction is assessed by the accuracy. It is calculated by dividing the total number of correct predictions by the total number of samples, as shown in Equation (4). Although it is a fundamental evaluation metric, it is insufficient in some cases, especially when the dataset is imbalanced.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

- Precision: Precision (P) is a performance metric that measures the proportion of correctly predicted positive instances among all instances predicted as positive. It is calculated as the ratio of true positives (TP) to the sum of true positives (TP) and false positives (FP), as shown in Equation (5).

$$P = \frac{TP}{TP + FP} \tag{5}$$

- Recall (Sensitivity): Recall (R) is a metric that measures the proportion of the actual positive instances that the model correctly identified. It is also known as Sensitivity or True Positive Rate (TPR). A high recall means that the model successfully identifies most of the true positives, which is critical in scenarios where missing a positive instance carries a significant cost, such as in disease detection. It is calculated as the ratio of true positives (TP) to the sum of true positives (TP) and false negatives (FN), as shown in Equation (6).

$$R = \frac{TP}{TP + FN} \tag{6}$$

- F1-Score: F1 score is the harmonic mean of precision and recall, see Equation (7). It is beneficial in cases of imbalanced datasets, where relying solely on accuracy, precision, or recall might give a misleading picture of model performance.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

4.2. Experimental Setup

All the experiments were performed using Google Colab [53] free GPU service. In the free service, the GPU memory was 15 GB, the System Ram was 12.7 GB, and the Disk Space was 112.6 GB. The model was trained using stratified K-fold cross-validation [57] with K = 3 to ensure a balanced class distribution across the folds. From each validation set, 50% of the data is further split into a test set. The 20 epochs, learning rate of 1e-5, and batch size of 32 were considered for training purposes, as shown in Table 2.

Table 2. Model Training Hyperparameters and Associated Values.

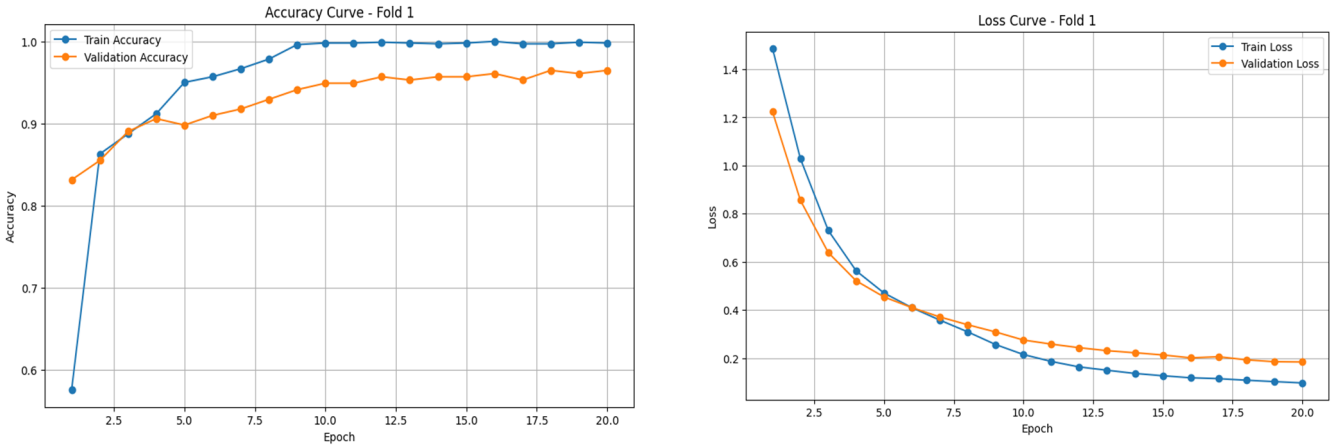
Training Hyperparameters	Value
Batch Size	32
Epochs	20
StratifiedKFold	3
Optimizer	Adam
Learning Rate	1×10^{-5}

4.3. Experimental Results

The model is trained for RGB images of 6 classes (3 real and three fake). The stratified K-fold with K = 3 approach is used to validate the model's generalizability in the real world. On each iteration, 50% of the data is separated from the validation set to serve as a test set for model testing. The images are resized to 224 × 224 and a patch size of 16 × 16. The total dataset size is 1532; each category has 265 images, except for the real medical class, which has 207 images.

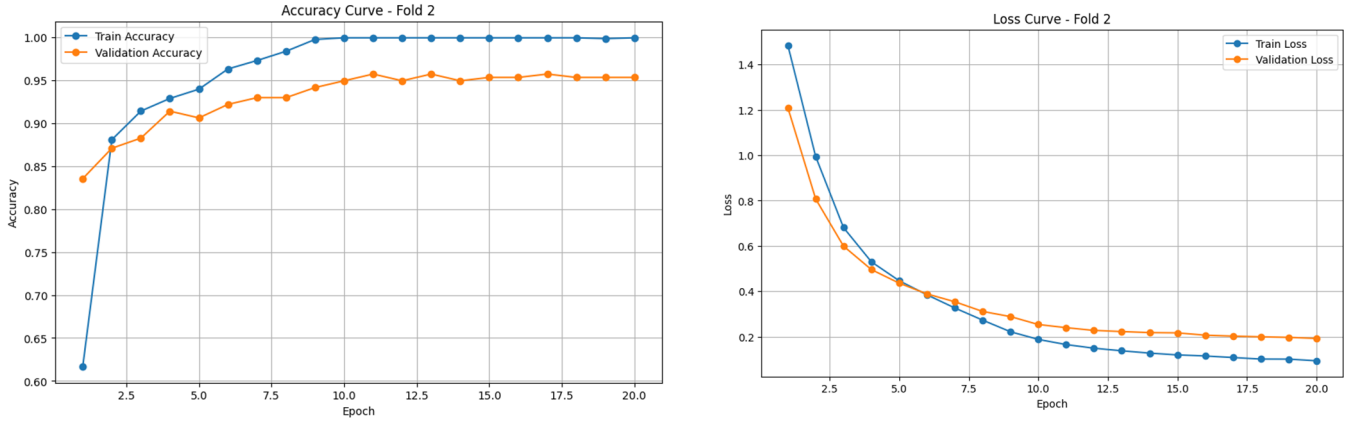
The learning curves for all three cross-validation folds are presented in Figures 9–11. Across all folds, the model demonstrates a consistent and stable training progression. The training accuracy smoothly converges to nearly perfect scores (99.80%, 99.90%, and 99.80% for

Folds 1–3, respectively), while the training loss steadily decreases to a value near zero. This indicates that the model effectively learned the underlying patterns from the training data.



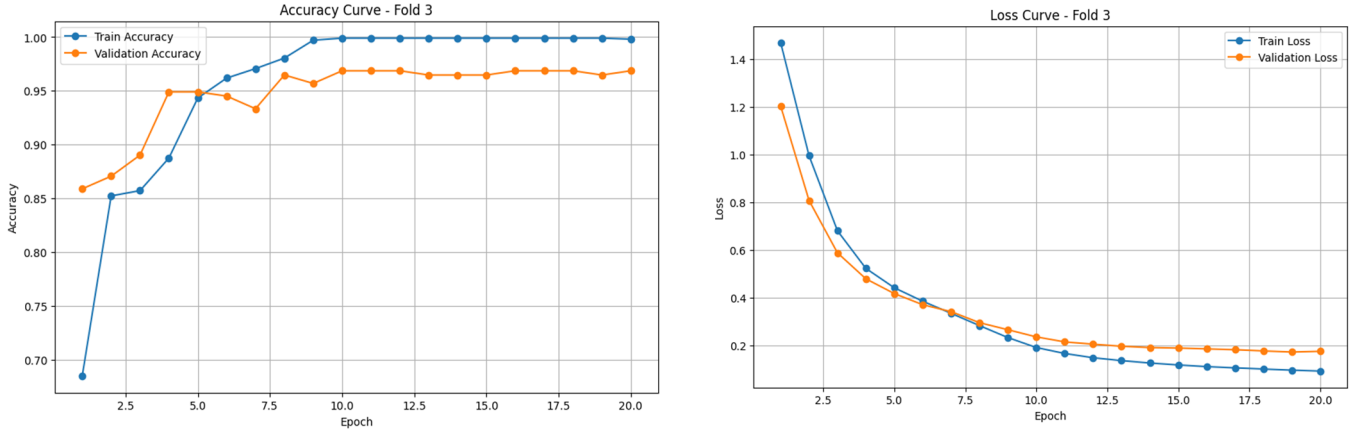
(**a**) (**b**)

Figure 9. Training and validation (**a**) accuracy and (**b**) loss curves for Fold 1 of the cross-validation.



(**a**) (**b**)

Figure 10. Training and validation (**a**) accuracy and (**b**) loss curves for Fold 2 of the cross-validation.



(**a**) (**b**)

Figure 11. Training and validation (**a**) accuracy and (**b**) loss curves for Fold 3 of the cross-validation.

Critically, the validation curves demonstrate strong generalization. The validation accuracy for each fold climbs rapidly and stabilizes at a high level of performance, reaching final values of 96.47% (Fold 1), 95.29% (Fold 2), and 96.86% (Fold 3). The corresponding validation loss for each fold decreases and plateaus at a low value (final losses of 0.1842, 0.1925, and 0.1745, respectively), without any signs of divergence. The minimal gap between the near-perfect training accuracies and these strong validation accuracies confirms that the model generalizes well and avoids significant overfitting. This consistent performance across all independent data folds underscores the robustness and stability of our proposed approach.

We have also considered the Receiver Operating Characteristics (ROC) curve to evaluate the classification performance of the proposed model. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different decision thresholds. The Area under the curve (AUC) provides a single metric to summarize this performance, where a higher value indicates better discriminative ability. To ensure reliability, ROC-AUC [58] curves were generated for each fold during cross-validation as shown in (Figure 12a–c).

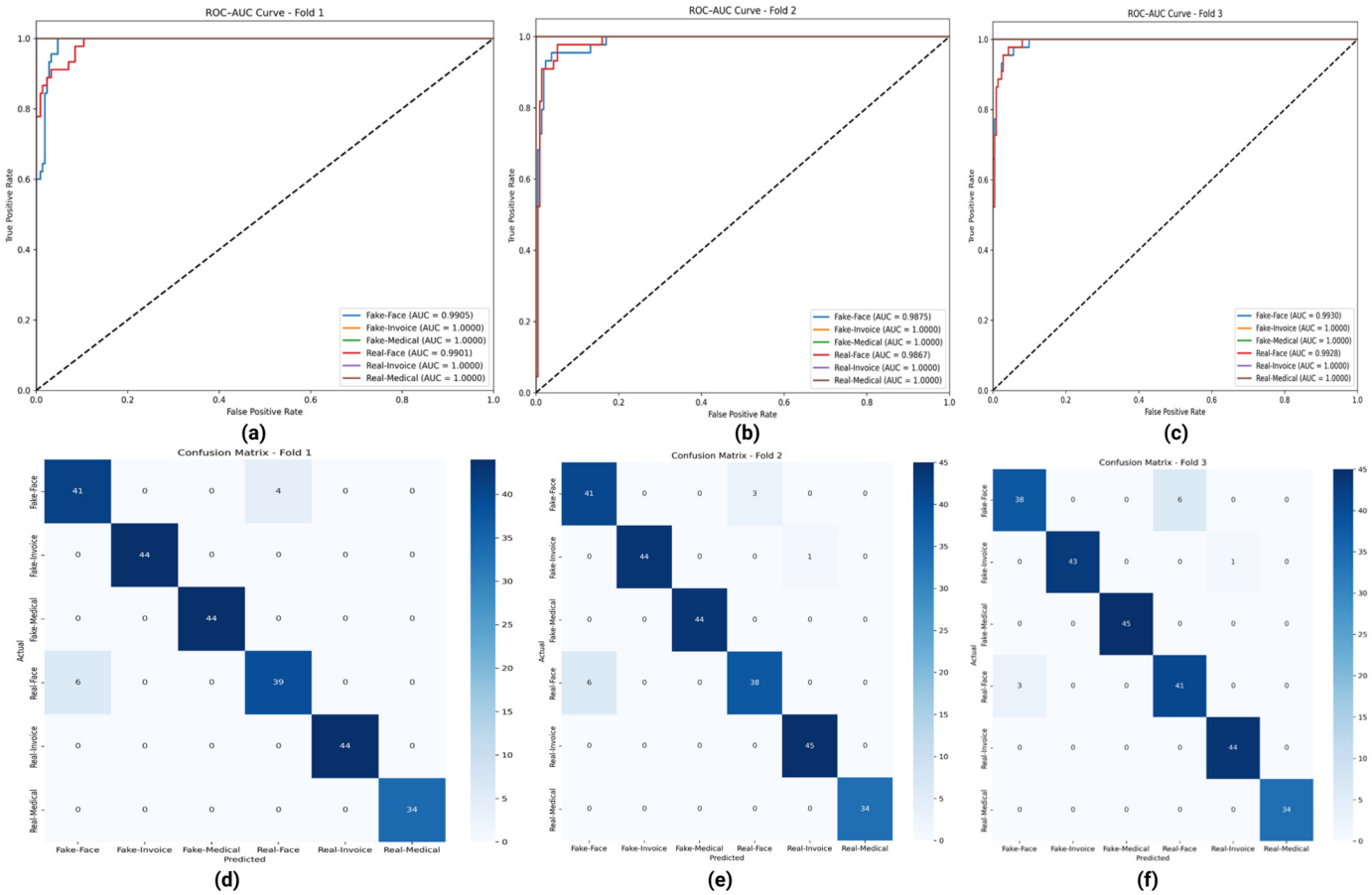


Figure 12. Performance evaluation across the 3-fold cross-validation. The top row displays the ROC-AUC curves for (**a**) Fold 1, (**b**) Fold 2, and (**c**) Fold 3. The bottom row displays the corresponding confusion matrices for (**d**) Fold 1, (**e**) Fold 2, and (**f**) Fold 3.

The results demonstrate an exceptionally high level of performance. Across all folds, the model achieved near-perfect AUC scores for all six classes. The majority of classes, particularly those in the invoice and medical domains, consistently reached a perfect AUC of 1.0000. The lowest score recorded across all tests was for the Real-Face class in Fold 2, which still achieved an outstanding AUC of 0.9867. These consistently high AUC

values, approaching the ideal of 1.0000, indicate that the model possesses an excellent capacity to distinguish between each specific class and all others. This result underscores the model's reliability in separating authentic from forged content across the three diverse domains.

To analyze the model's performance on a per-class basis, confusion matrices [59] were generated for the test set of each cross-validation fold (Figure 12d–f). The matrices reveal a consistently strong performance across all data splits, characterized by a dominant diagonal which indicates a high rate of correct predictions for every class. Notably, the model demonstrates near-perfect classification for both the invoice and medical domains. For these categories, there were minimal to zero misclassifications between authentic and fake samples across all three folds, highlighting the model's exceptional reliability in these structured-data contexts. The primary source of error consistently occurs between the Real-Face and Fake-Face classes. For instance, in Fold 1, six Real-Face images were misclassified as Fake-Face, and four Fake-Face images were misclassified as Real-Face. This specific confusion pattern is present in all folds and suggests that distinguishing between authentic and AI-generated faces is the most challenging task for the model. This finding aligns with the fact that modern facial forgery techniques are designed to be perceptually seamless, making them inherently more difficult to differentiate than forged documents or medical scans. The model's ability to still perform this difficult task with high accuracy further underscores the effectiveness of the ViT architecture.

Additionally, a detailed summary of the model's overall performance across the three cross-validation folds is presented in Table 3. The model demonstrates a high degree of stability and accuracy, achieving an average test accuracy of 96.08% and a weighted F1-score of 96.08%. The consistency of these metrics across the individual folds, with minimal variation, further validates the robustness of our proposed framework.

Table 3. Overall Model Performance Summary Scores.

	Training Accuracy	Validation Accuracy	Test Accuracy	Weighted Precision	Weighted Recall	Weighted F1
Fold 1	0.9980	0.9647	0.9609	0.9612	0.9609	0.9609
Fold 2	0.9990	0.9529	0.9609	0.9617	0.9609	0.9609
Fold 3	0.9980	0.9686	0.9608	0.9615	0.9608	0.9607
Average	0.9983	0.9620	0.9608	0.9614	0.9608	0.9608

The per-class classification reports for each fold (Tables 4–6) provide a more granular view of this performance and align perfectly with the insights from the confusion matrix analysis. The model exhibits exceptional performance for the invoice and medical domains, consistently achieving Precision, Recall, and F1-scores of or approaching 1.00. This is contrasted by the slightly lower, yet still strong, scores for the more challenging face categories. For example, in Fold 1 (Table 3), the F1-scores for Fake-Face and Real-Face were 0.8913 and 0.8864, respectively. This quantitative data corroborates our earlier finding that distinguishing between authentic and AI-generated faces represents the most significant challenge. The model's ability to maintain high overall accuracy, despite the difficulty of this specific task, underscores the effective and balanced performance of our cross-domain approach.

Table 4. Per Class performance as a classification report for the Fold-1 Test Set.

	Precision	Recall	F1	Support
Fake-Face	0.8723	0.9111	0.8913	45
Fake-Invoice	1.0000	1.0000	1.0000	44
Fake-Medical	1.0000	1.0000	1.0000	44
Real-Face	0.9070	0.8667	0.8864	45
Real-Invoice	1.0000	1.0000	1.0000	44
Real-Medical	1.0000	1.0000	1.0000	33
Accuracy			0.9609	256
macro avg	0.9632	0.9630	0.9629	256
weighted avg	0.9612	0.9609	0.9609	256

Table 5. Per Class performance as a classification report for the Fold-2 Test Set.

	Precision	Recall	F1	Support
Fake-Face	0.8723	0.9318	0.9011	44
Fake-Invoice	1.0000	0.9778	0.9888	45
Fake-Medical	1.0000	1.0000	1.0000	44
Real-Face	0.9268	0.8636	0.8941	44
Real-Invoice	0.9783	1.0000	0.9890	45
Real-Medical	1.0000	1.0000	1.0000	34
Accuracy			0.9609	256
macro avg	0.9629	0.9622	0.9622	256
weighted avg	0.9617	0.9609	0.9609	256

Table 6. Per Class performance as a classification report for the Fold-3 Test Set.

	Precision	Recall	F1	Support
Fake-Face	0.9268	0.8636	0.8941	44
Fake-Invoice	1.0000	0.9773	0.9885	44
Fake-Medical	1.0000	1.0000	1.0000	45
Real-Face	0.8723	0.9318	0.9011	44
Real-Invoice	0.9778	1.0000	0.9888	44
Real-Medical	1.0000	1.0000	1.0000	34
Accuracy			0.9608	256
macro avg	0.9628	0.9621	0.9621	256
weighted avg	0.9615	0.9608	0.9607	256

4.4. Comparsion with State of the Art CNN Based Pretrained Models

While a direct comparison to prior studies is challenging due to the novelty of our cross-domain approach, we benchmarked our proposed ViT model against a suite of established, pretrained CNN architectures to provide a robust performance context. The baseline models included DenseNet201 [50], MobileNetV2 [51], VGG19 [52], ResNet101V2 [60], and EfficientNetB0 [61]. To ensure a fair comparison, all models were initialized with ImageNet weights, and their classification heads were adapted for our task by adding dense layers with a dropout rate of 0.3. Each baseline was trained and evaluated using the same stratified 3-fold cross-validation methodology as our proposed model.

The results of this comparative analysis, presented in Tables 7 and 8, show a clear and significant performance gap. Our proposed ViT model substantially outperforms all CNN-based baselines. The strongest-performing CNN, DenseNet201, achieved a test accuracy of 91.01%, which our model surpassed by a margin of 5%. This considerable improvement highlights the architectural advantages of the transformer for this task.

Table 7. Overall Comparison with CNN-based Pretrained models.

	Average of 3-Folds					
	Training Accuracy	**Validation Accuracy**	**Test Accuracy**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**
VGG19	0.9986	0.9032	0.8996	0.9012	0.8996	0.8989
DenseNet201	0.9993	0.9163	0.9101	0.9119	0.9101	0.9097
MobileNetV2	0.9993	0.9137	0.8983	0.8987	0.8983	0.8974
ResNet101V2	0.9993	0.8980	0.9022	0.9042	0.9022	0.9018
EfficientNetB0	0.1733	0.1725	0.1721	0.0296	0.1721	0.0505
Proposed	0.9983	0.9620	0.9608	0.9614	0.9608	0.9608

Table 8. Average Per-Class Precision, Recall and F1 of 3-Folds for Proposed and CNN-Based Pretrained Models Comparison.

Models	**Metrics**	**Fake Face**	**Fake Invoice**	**Fake Medical**	**Real Face**	**Real Invoice**	**Real Medical**
VGG19	Precision	0.7508	1.0000	0.9852	0.7148	0.9781	1.0000
	Recall	0.6835	0.9774	1.0000	0.7741	1.0000	0.9804
	F1	0.7127	0.9885	0.9925	0.7413	0.9889	0.9901
DenseNet201	Precision	0.7579	1.0000	1.0000	0.7483	0.9858	1.0000
	Recall	0.7298	0.9852	1.0000	0.7663	1.0000	1.0000
	F1	0.7401	0.9924	1.0000	0.7539	0.9928	1.0000
MobileNetV2	Precision	0.7089	1.0000	1.0000	0.7141	0.9928	1.0000
	Recall	0.7141	0.9926	1.0000	0.7064	1.0000	1.0000
	F1	0.7105	0.9963	1.0000	0.7091	0.9963	1.0000
ResNet101V2	Precision	0.7211	1.0000	0.9779	0.7776	0.9709	1.0000
	Recall	0.7966	0.9699	1.0000	0.6919	1.0000	0.9706
	F1	0.7557	0.9847	0.9889	0.7305	0.9852	0.9851
EfficientNetB0	Precision	0.0575	0.0573	0.0573	0.0000	0.0000	0.0000
	Recall	0.3333	0.3333	0.3333	0.0000	0.0000	0.0000
	F1	0.0981	0.0978	0.0978	0.0000	0.0000	0.0000
Proposed	Precision	0.8905	0.9993	1.0000	0.9020	0.9854	1.0000
	Recall	0.9022	0.9850	1.0000	0.8874	1.0000	1.0000
	F1	0.8955	0.9924	1.0000	0.8939	0.9926	1.0000

A notable result is the failure of EfficientNetB0 to converge, achieving a test accuracy of only 17.21%. We hypothesize that its highly specialized architecture, which relies on a precise compound scaling principle, lacks the flexibility to learn generalizable features from such a heterogeneous dataset without extensive domain-specific fine-tuning. Overall, these results confirm the superiority of the patch-based transformer approach for cross-domain forgery detection. The ViT's self-attention mechanism is inherently better suited for capturing global contextual relationships and subtle, non-local artifacts characteristic of forgeries across diverse image types, a task where the more localized receptive fields of CNNs are less effective.

4.5. Practical Implications

This study demonstrates a generalizable framework for detecting AI-generated forgeries across multiple domains, including financial documents (invoices), human faces, and medical imagery. The findings have direct applications in insurance fraud prevention, identity verification, and safeguarding medical diagnostics, where the misuse of generative

models can lead to financial loss, reputational damage, or patient harm. By effectively handling heterogeneous data from both RGB and grayscale sources, the proposed method can be integrated into real-world automated screening systems to flag suspicious content before it reaches critical decision-making pipelines. Such a deployment can help organizations strengthen their digital forensic capabilities, improve trust in visual evidence, and maintain compliance with emerging regulatory requirements for synthetic media detection.

5. Conclusions

This study presented a framework for detecting GAN and diffusion-generated forgeries across heterogeneous domains, including human faces, invoices, and medical images. By standardizing preprocessing pipelines and utilizing a pretrained transformer architecture, our approach achieved consistently high performance, with test accuracies exceeding 0.96 (96%) across all categories. These results highlight the potential of ViT as a robust and scalable solution for cross-domain forgery detection. This approach addresses critical challenges in security-sensitive fields such as healthcare, finance, and identity verification.

Beyond demonstrating the feasibility of a single model for diverse forgery types, this work underscores the importance of input standardization when dealing with mixed RGB and grayscale datasets. Future work will focus on several key areas: scaling the framework to larger and more diverse datasets; evaluating the impact of higher input resolutions (e.g., 1024×1024) on detecting fine-grained forgeries, contingent on access to more powerful computational resources; assessing robustness against adversarial attacks; and extending detection capabilities to multimodal synthetic content. These advancements will pave the way for more resilient and trustworthy AI-driven forgery detection systems.

Author Contributions: Conceptualization, M.A.M., M.A.A. and A.M.; Methodology, M.A.M., M.A.A. and A.M.; Validation, M.A.M., M.A.A. and A.M.; Formal analysis, M.A.M., M.A.A. and A.M.; Investigation, M.A.M., M.A.A. and A.M.; Resources, M.A.M., M.A.A. and A.M.; Data curation, M.A.M., M.A.A. and A.M.; Writing—original draft, M.A.M., M.A.A. and A.M.; Writing—review and editing, M.A.M., M.A.A. and A.M.; Visualization, M.A.M., M.A.A. and A.M. 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.

Data Availability Statement: The dataset used in this study consists of six classes. Subsets of samples for five classes were obtained from publicly available sources (Kaggle), as cited in the manuscript. The authors generated samples of one class (Fake-Invoice), which are not publicly available due to ongoing related research. The data presented in this study are available on request from the corresponding author.

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

References

1. Rashid, A.B.; Kausik, M.A.K. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Adv.* **2024**, *7*, 100277. [CrossRef]
2. Friedrichs, N. Trusted Criminals. 2009. Available online: https://www.justicediwan.org/userfiles/David%20O_%20Friedrichs-Trusted%20Criminals_%20White%20Collar%20Crime%20in%20Contemporary%20Society%20%20-Wadsworth%20Pub%20Co%20(1995)(1).pdf (accessed on 10 April 2025).
3. Abbas, F.; Taeihagh, A. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Syst. Appl.* **2024**, *252*, 124260. [CrossRef]
4. Radiation Risk from Medical Imaging—Harvard Health. Available online: https://www.health.harvard.edu/cancer/radiation-risk-from-medical-imaging (accessed on 28 April 2024).

5. Solaiyappan, S.; Wen, Y. Machine learning based medical image deepfake detection: A comparative study. *Mach. Learn. Appl.* **2022**, *8*, 100298. [CrossRef]
6. Vecchietti, G.; Liyanaarachchi, G.; Viglia, G. Managing deepfakes with artificial intelligence: Introducing the business privacy calculus. *J. Bus. Res.* **2025**, *186*, 115010. [CrossRef]
7. The Worst Invoice Fraud Cases. Available online: https://xelix.com/the-worst-invoice-fraud-cases-in-2019-2020/ (accessed on 11 April 2025).
8. Ex-Employee Jailed After Defrauding National Trust Out of More Than £1 Million | The Crown Prosecution Service. Available online: https://www.cps.gov.uk/cps/news/ex-employee-jailed-after-defrauding-national-trust-out-more-ps1-million?utm_source=chatgpt.com (accessed on 18 September 2025).
9. ECJ: VAT Liability of Employee Issuing Fake Invoices in the Name of Employer/Tax & Legal—The Blog on Current Developments and Relevant Innovations/PwC Deutschland. Available online: https://blogs.pwc.de/en/german-tax-and-legal-news/article/241681/ecj-vat-liability-of-employee-issuing-fake-invoices-in-the-name-of-employer/ (accessed on 18 September 2025).
10. City of Fort Lauderdale Falls Victim to Phishing Scam Losing $1.2 Million—NBC 6 South Florida. Available online: https://www.nbcmiami.com/news/local/city-of-fort-lauderdale-falls-victim-to-phishing-scam-losing-1-2-million/3117167/ (accessed on 18 September 2025).
11. Zanardelli, M.; Guerrini, F.; Leonardi, R.; Adami, N. Image forgery detection: A survey of recent deep-learning approaches. *Multimed. Tools Appl.* **2022**, *82*, 17521–17566. [CrossRef]
12. Shinde, V.; Dhanawat, V.; Almogren, A.; Biswas, A.; Bilal, M.; Naqvi, R.A.; Rehman, A.U. Copy-Move Forgery Detection Technique Using Graph Convolutional Networks Feature Extraction. *IEEE Access* **2024**, *12*, 121675–121687. [CrossRef]
13. Amerini, I.; Ballan, L.; Caldelli, R.; Del Bimbo, A.; Serra, G. A SIFT-based forensic method for copy-move attack detection and transformation recovery. *IEEE Trans. Inf. Forensics Secur.* **2011**, *6*, 1099–1110. [CrossRef]
14. Yang, Z.; Liu, B.; Bi, X.; Xiao, B.; Li, W.; Wang, G.; Gao, X. D-Net: A dual-encoder network for image splicing forgery detection and localization. *Pattern Recognit.* **2024**, *155*, 110727. [CrossRef]
15. Li, Y.; Hu, L.; Dong, L.; Wu, H.; Tian, J.; Zhou, J.; Li, X. Transformer-Based Image Inpainting Detection via Label Decoupling and Constrained Adversarial Training. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 1857–1872. [CrossRef]
16. Rosberg, F.; Aksoy, E.E.; Alonso-Fernandez, F.; Englund, C. FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023.
17. Fijačko, N.; Štiglic, G.; Topaz, M.; Greif, R. Using OpenAI's Text-to-video Model Sora to Generate Cardiopulmonary Resuscitation Content. *Resuscitation* **2025**, *207*, 110484. [CrossRef]
18. Temsah, M.-H.; Nazer, R.; Altamimi, I.; Aldekhyyel, R.; Jamal, A.; Almansour, M.; Aljamaan, F.; Alhasan, K.; Temsah, A.A.; Al-Eyadhy, A.; et al. OpenAI's Sora and Google's Veo 2 in Action: A Narrative Review of Artificial Intelligence-driven Video Generation Models Transforming Healthcare. *Cureus* **2025**, *17*, e77593. [CrossRef]
19. Alanazi, S.; Asif, S. Understanding Deepfakes: A Comprehensive Analysis of Creation, Generation, and Detection. *Artif. Intell. Soc. Comput.* **2023**, *72*. [CrossRef]
20. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. proceedings.neurips.cc. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf (accessed on 11 July 2023).
21. Sharma, P.; Kumar, M.; Sharma, H.K.; Biju, S.M. Generative adversarial networks (GANs): Introduction, Taxonomy, Variants, Limitations, and Applications. *Multimed. Tools Appl.* **2024**, *83*, 88811–88858. [CrossRef]
22. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10674–10685. [CrossRef]
23. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. *Proc. Mach. Learn. Res.* **2021**, *139*, 8821–8831.
24. Introducing 4o Image Generation | OpenAI. Available online: https://openai.com/index/introducing-4o-image-generation/ (accessed on 25 July 2025).
25. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **2018**, *321*, 321–331. [CrossRef]
26. Rana, M.S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake Detection: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 25494–25513. [CrossRef]
27. Amiri, E.; Mosallanejad, A.; Sheikhahmadi, A. The Optimal Model for Copy-Move Forgery Detection in Medical Images. *J. Med. Signals Sens.* **2024**, *14*, 5. [CrossRef]
28. High Resolution Images Create a Pseudo-Pulmonary Embolism (PE) Type Appearance-Chest Case Studies-CTisus CT Scanning. Available online: https://www.ctisus.com/teachingfiles/cases/chest/285194 (accessed on 10 August 2024).

29. Albahli, S.; Nawaz, M. MedNet: Medical deepfakes detection using an improved deep learning approach. *Multimed. Tools Appl.* **2024**, *83*, 48357–48375. [CrossRef]
30. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. *Proc. Mach. Learn. Res.* **2021**, *139*, 10096–10106.
31. Akhtar, Z.; Dasgupta, D. A Comparative Evaluation of Local Feature Descriptors for Deepfakes Detection. ieeexplore.ieee.org. 2019. Available online: https://ieeexplore.ieee.org/abstract/document/9033005/ (accessed on 11 July 2023).
32. Sharafudeen, M.; Vinod Chandra, S.S. Medical Deepfake Detection using 3-Dimensional Neural Learning. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Swizerland, 2023; Volume 13739, pp. 169–180. [CrossRef]
33. Budhiraja, R.; Kumar, M.; Das, M.K.; Bafila, A.S.; Singh, S. MeDiFakeD: Medical Deepfake Detection using Convolutional Reservoir Networks. In Proceedings of the 2022 IEEE Global Conference on Computing, Power and Communication Technologies, GlobConPT 2022, New Delhi, India, 23–25 September 2022. [CrossRef]
34. Sharafudeen, M.; Chandra SS, V. Leveraging Vision Attention Transformers for Detection of Artificially Synthesized Dermoscopic Lesion Deepfakes Using Derm-CGAN. *Diagnostics* **2023**, *13*, 825. [CrossRef] [PubMed]
35. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784. [CrossRef]
36. Zhang, J.; Huang, X.; Liu, Y.; Han, Y.; Xiang, Z. GAN-based medical image small region forgery detection via a two-stage cascade framework. *PLoS ONE* **2024**, *19*, e0290303. [CrossRef]
37. Arshed, M.A.; Mumtaz, S.; Gherghina, Ș.C.; Urooj, N.; Ahmed, S.; Dewi, C. A Deep Learning Model for Detecting Fake Medical Images to Mitigate Financial Insurance Fraud. *Computation* **2024**, *12*, 173. [CrossRef]
38. Bekci, B.; Akhtar, Z.; Ekenel, H.K. Cross-Dataset Face Manipulation Detection. In Proceedings of the 2020 28th Signal Processing and Communications Applications Conference, SIU 2020-Proceedings, Gaziantep, Turkey, 5–7 October 2020. [CrossRef]
39. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3204–3213. [CrossRef]
40. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
41. Korshunov, P.; Marcel, S. DeepfakeTIMIT. Available online: https://www.idiap.ch/en/scientific-research/data/deepfaketimit (accessed on 22 September 2025).
42. Eyebrow Recognition for Identifying Deepfake Videos | IEEE Conference Publication | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/9211068/authors#authors (accessed on 12 July 2023).
43. GitHub-LeeDongYeun/Deepfake-Detection. Available online: https://github.com/LeeDongYeun/deepfake-detection (accessed on 27 May 2025).
44. Silva, S.H.; Bethany, M.; Votto, A.M.; Scarff, I.H.; Beebe, N.; Najafirad, P. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Sci. Int.* **2022**, *4*, 100217. [CrossRef]
45. High-Quality Invoice Images for OCR. Available online: https://www.kaggle.com/datasets/osamahosamabdellatif/high-quality-invoice-images-for-ocr?resource=download (accessed on 28 June 2025).
46. DALL·E 3 | OpenAI. Available online: https://openai.com/index/dall-e-3/ (accessed on 13 June 2025).
47. 140k Real and Fake Faces | Kaggle. Available online: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces (accessed on 12 July 2023).
48. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *43*, 4217–4228. [CrossRef]
49. BTD-MRI and CT Deepfake Test Sets. Available online: https://www.kaggle.com/datasets/freddiegraboski/btd-mri-and-ct-deepfake-test-sets (accessed on 25 July 2025).
50. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
51. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [CrossRef]
52. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
53. Rafique, R.; Gantassi, R.; Amin, R.; Frnda, J.; Mustapha, A.; Alshehri, A.H. Deep fake detection and classification using error-level analysis and deep learning. *Sci. Rep.* **2023**, *13*, 7422. [CrossRef]
54. Pareek, J.; Jacob, J. *Data Compression and Visualization Using PCA and T-SNE*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 135, pp. 327–337. [CrossRef]

55. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR 2021-9th International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
56. Arshed, M.A.; Alwadain, A.; Ali, R.F.; Mumtaz, S.; Ibrahim, M.; Muneer, A. Unmasking Deception: Empowering Deepfake Detection with Vision Transformer Network. *Mathematics* **2023**, *11*, 3710. [CrossRef]
57. Prusty, S.; Patnaik, S.; Dash, S.K. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front. Nanotechnol.* **2022**, *4*, 972421. [CrossRef]
58. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
59. Valero-Carreras, D.; Alcaraz, J.; Landete, M. Comparing two SVM models through different metrics based on the confusion matrix. *Comput. Oper. Res.* **2023**, *152*, 106131. [CrossRef]
60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7–13 December 2015; pp. 770–778. [CrossRef]
61. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; pp. 10691–10700.