computers

MDPI

Article

Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model

Muhammad Asad Arshed [1,*], Shahzad Mumtaz [2], Muhammad Ibrahim [2], Christine Dewi [3,*], Muhammad Tanveer [1] and Saeed Ahmed [1,4]

[1] School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan; muhammad_tanveer@umt.edu.pk (M.T.); saeed.ahmed@med.lu.se (S.A.)
[2] Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; shahzad.mumtaz@iub.edu.pk (S.M.); muhammad.ibrahim@iub.edu.pk (M.I.)
[3] Department of Information Technology, Satya Wacana Christian University, Salatiga 50715, Indonesia
[4] Department of Experimental Medical Science, Biomedical Center (BMC), Lund University, 22184 Lund, Sweden
* Correspondence: muahmmadasadarshed@gmail.com (M.A.A.); christine.dewi@uksw.edu (C.D.)

Abstract: In response to the rapid advancements in facial manipulation technologies, particularly facilitated by Generative Adversarial Networks (GANs) and Stable Diffusion-based methods, this paper explores the critical issue of deepfake content creation. The increasing accessibility of these tools necessitates robust detection methods to curb potential misuse. In this context, this paper investigates the potential of Vision Transformers (ViTs) for effective deepfake image detection, leveraging their capacity to extract global features. **Objective:** The primary goal of this study is to assess the viability of ViTs in detecting multiclass deepfake images compared to traditional Convolutional Neural Network (CNN)-based models. By framing the deepfake problem as a multiclass task, this research introduces a novel approach, considering the challenges posed by Stable Diffusion and StyleGAN2. The objective is to enhance understanding and efficacy in detecting manipulated content within a multiclass context. **Novelty:** This research distinguishes itself by approaching the deepfake detection problem as a multiclass task, introducing new challenges associated with Stable Diffusion and StyleGAN2. The study pioneers the exploration of ViTs in this domain, emphasizing their potential to extract global features for enhanced detection accuracy. The novelty lies in addressing the evolving landscape of deepfake creation and manipulation. **Results and Conclusion:** Through extensive experiments, the proposed method exhibits high effectiveness, achieving impressive detection accuracy, precision, and recall, and an F1 rate of 99.90% on a multiclass-prepared dataset. The results underscore the significant potential of ViTs in contributing to a more secure digital landscape by robustly addressing the challenges posed by deepfake content, particularly in the presence of Stable Diffusion and StyleGAN2. The proposed model outperformed when compared with state-of-the-art CNN-based models, i.e., ResNet-50 and VGG-16.

Keywords: deep learning; image processing; CNN; deepfake identification; artificial intelligence; stable diffusion; StyleGAN2; vision transformer; global feature extraction; patches

Citation: Arshed, M.A.; Mumtaz, S.; Ibrahim, M.; Dewi, C.; Tanveer, M.; Ahmed, S. Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model. *Computers* **2024**, *13*, 31. https://doi.org/10.3390/computers13010031

Academic Editor: Lucia Maddalena

Received: 29 November 2023
Revised: 17 January 2024
Accepted: 18 January 2024
Published: 21 January 2024

1. Introduction

Over the past decade, social media content, including photos and videos, has seen a remarkable surge driven by the widespread availability of affordable devices like smartphones, cameras, and computers. The proliferation of social media platforms has facilitated the swift sharing of such content, resulting in exponential growth of online material and easy accessibility for users [1].

Simultaneously, there have been significant advancements in machine learning (ML) and deep learning (DL) algorithms, which are highly efficient in manipulating audiovisual content [1]. Unfortunately, this technological progress has also created and disseminated

deepfakes, i.e., synthetic audio and video content generated using AI algorithms [2,3]. The rapid development of deepfake technology poses a serious threat [4] as it can be utilized to spread disinformation globally and potentially sway public opinion. In instances such as election manipulation or character defamation, the ease of spreading false information can be exploited.

As deepfake creation becomes more sophisticated, the authentication and verification of video evidence in legal disputes and criminal court cases could become increasingly challenging [5]. Ensuring the integrity and reliability of video submissions as evidence will demand significant scrutiny, particularly in the face of advanced deepfake techniques [6]. Moreover, the exponential growth of social media content and the evolution of deepfake technology raises concerns about the potential misuse and manipulation of information, demanding further attention from researchers, policymakers, and the technology community [7]. The production of high-resolution deepfake images relies on intricate algorithms commonly based on DL models like GANs. These complex DL techniques are crucial in creating realistic and convincing synthetic images [8].

The proliferation of deepfake technology gives rise to numerous concerns and potential dangers across various industries [9]. One significant area impacted is cybersecurity [10], where the ability to manipulate facial photos convincingly raises alarms about identity theft, deception, and unauthorized access to sensitive information. Moreover, the widespread use of deepfakes poses a substantial risk to public trust, as malicious individuals can exploit this technology to create deceitful visual cues, propagate misinformation, or tarnish the reputations of others [11]. Due to these issues, researchers and academics have been focusing on devising methods to detect and mitigate the adverse effects of deepfakes. By developing advanced approaches, they aim to safeguard individuals and organizations from the potential harms posed by this evolving technology [12]. This involves harnessing the progress made in computer vision, machine learning, and forensic analysis to detect crucial indicators of image manipulation and effectively differentiate between authentic and manipulated facial images [13].

Various approaches have been put forward to detect deepfakes, and a significant portion relies on deep learning techniques [14]. The United States Defense Advanced Research Projects Agency (DARPA) has initiated a media forensic research project to develop effective methods for detecting fake media [15]. This endeavor reflects the growing importance of addressing the challenges posed by deepfake technology in safeguarding the authenticity and credibility of digital media content [15]. Additionally, Facebook, in collaboration with Microsoft, has introduced an AI-based deepfake identification challenge. This joint effort signifies the industry's commitment to combatting the risks associated with deepfake technology by fostering the development of advanced AI solutions for detecting and countering deceptive media content [16].

Recently, numerous prominent techniques have been put forward for identifying fake images. However, these models often exhibit limited generalization capability, leading to a drop in performance when faced with the latest deepfake or manipulation methods. Akhtar et al. [17] considered Convolutional Neural Network (CNN)-based SqueezeNet [18], VGG16 [19], ResNet [20], DenseNet [21], and GoogleNet [22] in their study for the identification of face manipulation. The models demonstrated impressive accuracy when tested on the same manipulation type they were trained on. However, their performance declined when confronted with novel manipulations not part of their training dataset. To address the issues mentioned above, this study adopts the Vision Transformer (ViT) model. The input image is divided into blocks during the general training process, treating each block as a separate entity. The ViT employs self-attention modules to understand the relationships between these embedded patches. The ViT has demonstrated exceptional performance in standard classification tasks by emphasizing important features while reducing the impact of noisy ones through its self-attention mechanism. Inspired by this perspective, this study proposes a deepfake image identification network based on the ViT. The experimental

results indicate that the proposed network achieves satisfactory outcomes in deepfake image detection. This research contributes to the field in the following ways:

- Our primary contribution lies in being the first to address this problem as a multi-classification task. No prior work has tackled this specific aspect, and our study represents a pioneering effort in this area. By approaching deepfake detection through the lens of multi-classification, we aim to enhance the accuracy and efficacy of identifying and categorizing deepfake content, thereby advancing the field's understanding and capabilities in combating this evolving challenge.
- We have compiled and curated our dataset specifically for multiclass deepfake identification. This dataset is carefully designed to facilitate the training and evaluation of our deepfake detection model, allowing us to explore the complexities of multiclass classification and improve the accuracy of deepfake identification.
- The proposed fine-tuned ViT model exhibits superior performance to state-of-the-art deepfake identification models.
- Following an extensive analysis, our research firmly establishes the remarkable robustness and generalizability of the proposed method, surpassing numerous state-of-the-art techniques. The findings validate the effectiveness and reliability of our approach in the field of deepfake detection.

The remainder of this paper is divided as follows. Section 2 provides the survey's existing methods, emphasizing the role of the ViT. Section 3 outlines the methodology of the ViT's application, while the experimental results showcase its effectiveness. The discussion interprets findings and outlines future implications for multimedia forensics in Section 4, and Section 5 provides the conclusion of this study.

2. Related Works

The proliferation of deepfake technology has ushered in a new era of challenges in the realm of multimedia forensics and information veracity. Prior research has underscored the need for innovative methods to detect and combat the manipulation of digital content [23]. Early efforts in deepfake detection centered around traditional signal processing and image analysis techniques. Researchers leveraged facial landmarks, inconsistencies in lighting, and unnatural facial movements as indicators of potential manipulation. However, the rapid advancement of GANs led to the creation of more convincing and challenging-to-detect deepfakes, necessitating a shift towards more sophisticated detection methods. Akhtar and Dasgupta [24] investigated the feasibility of utilizing local feature descriptors to recognize manipulated faces. Their study presented a comparative experimental analysis of ten local feature descriptors, employing the 'DeepfakeTIMIT' database as a testing ground.

Bekci et al. [25] presented a deepfake detection system that leverages metric learning and steganalysis-rich models to enhance performance against unseen data and manipulations. To evaluate the effectiveness of their approach, an empirical analysis was conducted using openly accessible datasets, including FaceForensics++, DeepFakeTIMIT, and CelebDF. The suggested framework demonstrated significant accuracy improvements ranging from 5% to 15% when faced with concealed modifications. Li et al. [26] investigated the differences in eye-blinking patterns between deepfake videos and those displayed by genuine human subjects. Based on their observations, they developed a novel eye-blinking detection technique tailored to identify deepfake videos specifically.

In their study, Nguyen et al. [27] used the eyebrow region as a set of features to identify deepfake videos. They applied four deep learning methods—LightCNN, Resnet, DenseNet, and SqueezeNet—for this purpose. The UADFV and Celeb-DF datasets produced the highest AUC (Area Under Curve) values of 0.984 and 0.712, respectively.

Patel et al. [28] introduced Trans-DF, a deepfake detection method relying on random forests. The Trans-DF model demonstrated impressive detection accuracy, achieving a high score of 0.902, highlighting its effectiveness in identifying deepfake videos. Another approach was presented by Yang and colleagues, utilizing SVM classifiers to differentiate between deepfake images and videos. Their method capitalized on variations in head poses

as essential features for discrimination. Through the implementation of this technique, they created a system with a noteworthy AUROC score of 0.890, effectively detecting and distinguishing deepfake content.

Ciftci et al. [29] presented a pioneering technique to trace the origins of deepfake content by scrutinizing biological cues within residuals. This groundbreaking study marked the inaugural application of biological indicators in the detection of deepfake sources. The researchers performed experimental assessments on the Face Forensics++ dataset, incorporating numerous ablation tests to affirm the validity of their method. Notably, they attained a remarkable accuracy rate of 93.39% in source identification across four distinct deepfake generators. These results emphasize the efficacy of their proposed approach and its promising ability to accurately trace the roots of deepfake content.

In 2022, Yang et al. [30] introduced a deepfake detection model named MSTA_Net, leveraging machine learning techniques. This model specifically examined the texture properties of an image to discern abnormalities indicative of deepfake alterations. Unlike other approaches that focused solely on facial regions, the MSTA_Net model considered the entire image. By establishing connections between manipulated and unmanipulated areas within the image, the model identified irregularities in texture and signaling variations as potentially fake. Conversely, when no irregularities were detected, the image received a non-fake label, suggesting a higher likelihood of authenticity. Their proposed model facilitated the identification of genuine and manipulated images based on their overall texture characteristics. In recent studies, the prominence of multi-attentional and transformer models has grown significantly in the area of deepfake detection [31]. Overall, the multi-modal, multi-scale transformer model presented by Wang et al. [32] offers a promising approach to deepfake detection. By enabling the analysis of image patches at different spatial levels and utilizing multiple modalities, the model aims to improve accuracy and robustness in identifying deepfake content.

CNNs have demonstrated remarkable efficacy in detecting deepfake content, underscoring their importance in this field. Despite their proficiency in extracting features from small objects, CNNs may encounter challenges in precisely identifying key regions within an image. Leveraging a ViT model for deepfake identification presents an intriguing and promising alternative. ViTs were originally introduced for image classification tasks and have demonstrated strong performance on various computer vision benchmarks [33]. There are many reasons to choose ViTs for this study, of which the main ones are listed below.

- Attention Mechanism: ViT models utilize self-attention mechanisms, which allow them to capture long-range dependencies within an image. This is crucial for detecting subtle inconsistencies and artifacts that might be present in deepfake images. Deepfake generation often involves stitching or blending different parts of images, and attention mechanisms can help identify these anomalies.
- Global Context: Classic CNNs are great at pulling out details from specific areas, whereas ViT models take in the complete image as a sequence of patches, allowing them to grasp the global context. This difference can be beneficial for deepfake detection, as it lets the model scrutinize the overall structure and consistency of an image.
- Robustness to Manipulations: ViT models might exhibit increased robustness to common manipulation techniques used in deepfake generation. Their attention mechanisms can potentially make them more resistant to simple modifications like noise addition or small alterations in pixel values.
- Interpretable Attention Maps: ViT models generate attention maps that indicate which parts of an image are considered the most important for making predictions. These maps could provide insights into how the model distinguishes between real and deepfake images, aiding in understanding and improving the model's decision-making process.

3. Proposed Methodology

This section outlines and presents the methodologies utilized and proposed to identify fake images accurately. These methods are carefully designed to enhance the precision and effectiveness of detecting and distinguishing fake content from genuine ones.

3.1. Dataset

For our experiment, we utilized a dataset sourced from Kaggle [34], an online source [35], Stable Diffusion [36], and the StyleGAN2 encoding of Stable Diffusion [37]. We used the free version of TPU (Tensor Processing Unit) that is provided by Google Colab to prepare the dataset as well as for research experiments.

1. **Real Images:** We considered Kaggle [34] for real images; due to the limitation of computation power, we considered 10K images from this source.
2. **Online Source:** We obtained GAN-based fake images from an online source [35]. This source consistently provides new fake images with each visit, enabling us to access a diverse and up-to-date dataset for our analysis and experimentation.
3. **Stable Diffusion:** In this study, we curated a dataset focused on Stable Diffusion, specifically in the context of text-to-image conversion. Stable Diffusion text-to-image conversion involves a method for consistently generating high-quality images from textual descriptions. The primary objective is to create realistic and cohesive images that faithfully represent the provided textual descriptions. This approach utilizes advanced machine learning models and deep learning techniques to achieve this goal. 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. Techniques such as word embeddings or attention mechanisms may be employed to capture the semantic meaning of the text. Following this, the image synthesis model utilizes the encoded text to produce a corresponding image, as illustrated in Figure 1. The image synthesis process is geared towards capturing the visual details and context outlined in the text description. To ensure stability and consistency in the image generation process, regularization techniques and control mechanisms may be incorporated. Stable Diffusion text-to-image conversion has various applications, including creative content generation, virtual world creation, and multimedia production. As this technology continues to advance, the generation of fake content and the potential for misuse of such tools are steadily increasing. This trend poses significant challenges and concerns in various domains, such as disinformation campaigns, image manipulation, and privacy breaches. Stable Diffusion based on the conditional Latent Diffusion Model (L_{DM}) and the equation of L_{DM} concerning conditional image pairs can be seen in Equation (1) [36]. In Equation 1, models can be understood as a series of equally weighted denoising autoencoders, denoted as $\varepsilon_\vartheta(x_t, t)$ for t = 1...T. These autoencoders are trained to predict a denoised version of their input, where x_t represents a noisy version of the input x.
4. **StyleGAN2 encoding of Stable Diffusion:** This dataset is available on Kaggle [37] with the name *Synthetic Faces High Quality (SFHQ)*. This dataset comprises high-quality 1024 × 1024 curated face images. It was created through a multi-step process. Firstly, a significant number of "text to image" generations were generated, primarily using Stable Diffusion v2.1, along with some from Stable Diffusion v1.4 models. Subsequently, a set of photo-realistic candidate images was generated by encoding these images into the latent space of StyleGAN2 and applying a small manipulation to enhance each image into a high-quality, photo-realistic candidate. This process ensured that the dataset contained diverse and visually appealing face images, enabling us to conduct comprehensive and accurate analyses in our research. The styleGAN2 is mathematically based on a generator network (G), mapping vector (F), noise vector (z), conditional vector (y), and style vector (s) to produce the synthesized image; see Equation (2) that is used to synthesize the image x.

$$L_{DM} = E_{x,\varepsilon\sim N(0,1),t}\left[||\varepsilon - \varepsilon_\vartheta(x_t, t)||_2^2\right] \quad (1)$$

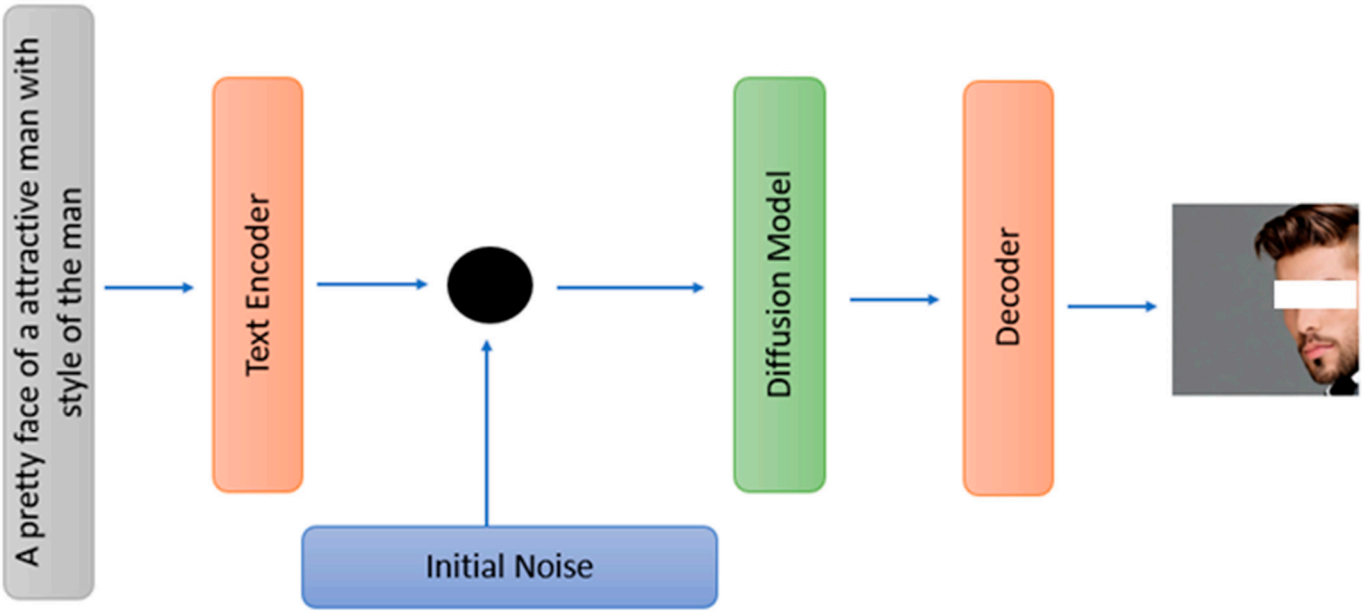


Figure 1. Sample diagram of text-to-image generation with stable diffusion.

$$x = G(z, y, s) \quad (2)$$

The style vector (s) is computed with mapping network (F) with Equation (3).

$$s = F(z) \quad (3)$$

In the context of StyleGAN2, the generator G and the mapping network F are trained to generate high-quality images by considering the style information (s) along with noise (z) and conditioning (y) inputs.

In our research, we have ultimately focused on four distinct classes and taken the initiative to address the deepfake detection problem using a multiclass approach. By considering multiple classes (Real: 10,000, GAN_Fake: 10,000, Diffusion_Fake: 10,000, and Stable&Gan_Fake: 10,000), we aim to enhance the precision and reliability of our deepfake detection model, accommodating a broader range of deepfake variations and increasing its potential for real-world applications.

To overcome the challenge of class imbalance and potential model bias, we meticulously prepared the dataset in a balanced format. By ensuring each class has a similar representation, we aim to create a more equitable training environment for our deepfake detection model. This approach helps mitigate the impact of overrepresented or underrepresented classes, leading to a fairer and more robust model capable of accurately identifying deepfake content across all classes. Sample images from the prepared dataset can be found in Table 1.

Table 1. Prepared dataset images.

Real	GAN_Fake	Diffusion Fake	Stable&GAN Fake
			

3.2. ViT Architecture

In this section, we introduce the ViT framework, delving into its core principles, structure, self-attention mechanism, multi-headed self-attention, and the mathematical foundations that shape its design. The ViT emerged in 2020 [38] as a groundbreaking paradigm in computer vision, revealing its potential to redefine our approach to image analysis and comprehension. Initially rooted in the Transformer architecture crafted for natural language processing, the ViT introduces a novel concept by treating images as sequences of tokens, commonly represented by image patches. With the transformer design, ViT adeptly processes these token sequences, enabling effective image analysis and understanding in a sequence-based manner.

A key strength of ViT lies in its adaptability and versatility. The foundational transformer architecture has demonstrated remarkable success across diverse tasks, including picture restoration and object detection. This underscores the broad applicability and effectiveness of the ViT framework, positioning it as a potent tool in the field of computer vision with the potential to revolutionize our approach to image-related tasks [39].

Tokenization and embedding stand as crucial steps within the ViT architecture. When handling the input image, it undergoes initial division into a grid of non-overlapping patches. Subsequently, these patches are flattened and transformed into a higher-dimensional space through a linear operation, followed by normalization. This method endows the ViT model with the capability to capture both global and local information from the image, promoting comprehensive learning. It enables the model to effectively grasp the intricate features and context of the image. The synergy between tokenization and embedding plays a pivotal role in empowering ViT to excel in a variety of computer vision tasks.

The ViT architecture can be mathematically represented by assuming X is a set of image patches extracted from the input image. Each patch is a vector representing a portion of the image. The set of patches (X) is represented in Equation (4), where N is the number of patches.

$$X = \{x_1, x_2, x_3, \ldots., x_N\} \tag{4}$$

The ViT model consists of several components that are enlisted below (also see Figure 2).

- **Patch Embedding:** The image patches ($x_1, x_2 \ldots x_N$) are linearly projected to an embedding space by a linear transformation W_{patch} (see Equation (5)).

$$E = \left\{W_{patch}x_1, W_{patch}x_2, \ldots, W_{patch}x_N\right\} \tag{5}$$

- **Positional Embedding:** Each patch embedding (e_1, e_2, $\ldots.e_N$) is augmented with positional information (p_1, $p_2 \ldots..p_n$) to capture spatial relationships. These positional embeddings are added to the patch embeddings (see Equation (6)).

$$E_{POS} = \{e_1 + p_1, e_2 + p_2 \ldots\ldots e_N + +p_N\} \tag{6}$$

- **Transformer Encoder:** The transformer encoder processes the positional embeddings E_{pos}. This encoder comprises several layers, each incorporating self-attention

mechanisms and feedforward neural networks. The result of this encoding is a collection of contextualized embeddings, as depicted in Equation (6). Equation (7), $(z_1, z_2, \ldots\ldots, z_N)$, represents the output representations or embeddings produced by the Transformer encoder for each position in the input sequence.

$$TransformerEncoder(E_{POS}) = \{z_1, z_2, \ldots\ldots, z_N\} \quad (7)$$

- **Classification Head:** The final contextualized embeddings Z are used for downstream tasks. In classification tasks, a classification head takes the average or a specific token's embedding (e.g., classification token) from Z and passes it through one or more fully connected layers to make predictions.

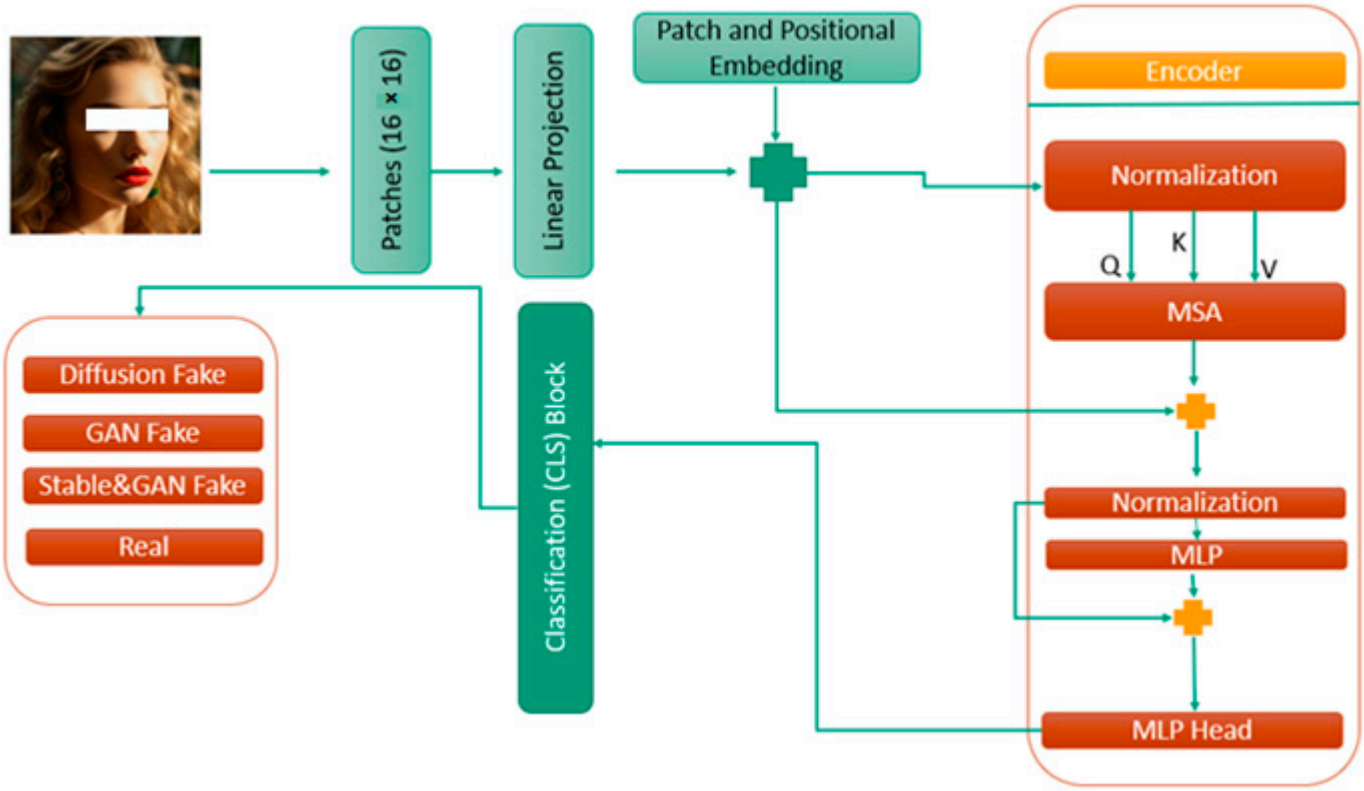


Figure 2. Model base architecture diagram [38].

The ViT design centers around the Multi-head Self-Attention (*MSA*) mechanism, which plays a pivotal role in the model's capabilities. MSA empowers the ViT to attend to multiple parts of the image simultaneously. It consists of distinct "heads", with each head independently computing attention. By focusing on different regions of the image, these attention heads produce various representations, which are then concatenated to generate the final image representation. This approach enables the ViT to capture intricate interactions between input elements by attending to multiple sections simultaneously. However, this enhancement comes at the cost of increased complexity and computational requirements. The utilization of multiple attention heads and the subsequent aggregation of their outputs necessitate more computational resources. The mathematical representation of *MSA* can be seen in Equation (8).

$$MSA(Q, K, V) = Concat\ (H1, H2, \ldots, Hn) \quad (8)$$

In Equation (7), Q, K, and V stand for the query, key, and value matrices, respectively. The H_1, H_2,... H_n represents the output of multiple attention heads. In the context of neural networks, particularly in transformers, a multi-head attention mechanism involves

using multiple sets of attention weights (attention heads) to capture different aspects of relationships in the input data. Each H_i is the output of the i-th attention head. The self-attention mechanism plays a pivotal role in transformers, serving as the foundational component for explicitly modeling interactions and relationships across all sequences in prediction tasks. Unlike CNNs, which depend on local receptive fields, the self-attention layer gathers insights and features from the entire input sequence, allowing it to capture both local and global information. This unique characteristic distinguishes self-attention from CNNs, as it promotes a more comprehensive interpretation and representation of information, leading to improved performance in various sequence-based tasks.

The attention mechanism involves computing the dot product between the query and key vectors, followed by normalization using SoftMax. Subsequently, it modulates the value vectors to generate an enhanced output representation, a task carried out in the CLS block. Figure 2 is the base abstract architectural diagram of the ViT model [38].

3.3. ViT Hyper-Parameters

In this study, the initial images undergo preprocessing and are divided into patches measuring 16×16 pixels, subsequently scaled to 224×224 pixels. This reduction technique involves breaking down the image into smaller fixed-size patches, each with dimensions of 16 pixels in width and 16 pixels in height.

The model employed in this study underwent training on a substantial dataset known as ImageNet-21k. This dataset encompasses around 14 million photos, categorized into 21,841 distinct classes, making it specifically tailored for extensive image classification tasks. The model's architecture comprises 12 transformer layers, each housing 768 hidden components. Its overall capacity is reflected in its 85.8 million trainable parameters, which play a significant role in the learning process. For a comprehensive understanding, the values and configurations of the parameters used in the ViT model are detailed in Table 2.

Table 2. Parameter configurations.

Parameters	Values
Transformer Encoder Hidden Layers	12
Hidden Layer Activation	Gelu
Channels	3
Patches	16×16
Balanced	True
Learning Rate	2×10^{-5}
Epochs	5

Figure 3 showcases the abstract-level diagram illustrating the proposed methodology. This diagram provides an overview of the key components and steps involved (dataset preparation, preprocessing, splitting, model tuning, training, and evaluation) in our approach, offering a visual representation of how our method operates and achieves its objectives.

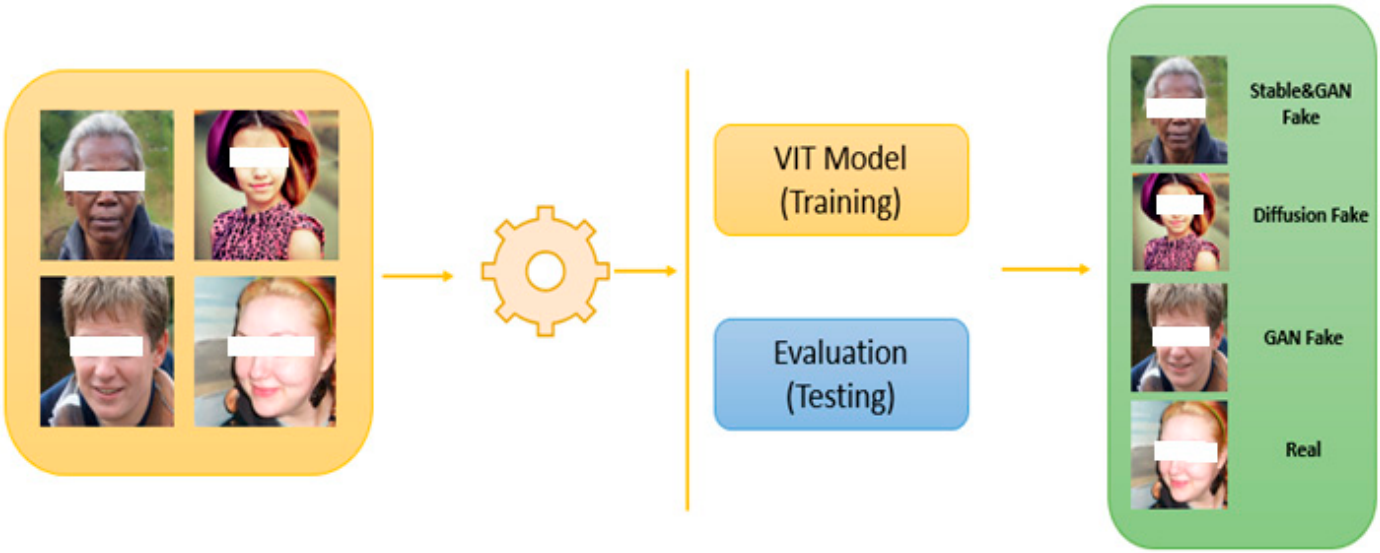


Figure 3. Abstract-level diagram of the proposed methodology.

3.4. CNN Architecture-Based Pretrained Models

The primary objective of this study was to uncover and identify the most recently manipulated deepfake images, specifically those generated using Stable Diffusion and StyleGAN2. This research stands out as a pioneering effort not only in recognizing these cutting-edge manipulated fake images but also in addressing the challenge in a multiclass context.

To demonstrate the effectiveness of patch technology over traditional CNN and CNN-based pretrained models such as VGG16 and ResNet50, this study employed a fine-tuning approach. The models were preloaded with weights from the ImageNet dataset using a weight transfer technique. In this process, the network layers were frozen, and the last fully connected layers were omitted from the architectures.

To adapt these models for our purposes, a flattened layer was introduced to eliminate the fully connected layers, and dense layers with four neurons were added. The activation function was set to SoftMax to tackle the multiclass nature of the problem. This nuanced approach aims to showcase that, in the realm of manipulated deepfake image detection, patch technology can outperform the more conventional CNN and pretrained models. The local feature extraction is the main reason for selecting CNN-based models.

4. Experiment Results and Discussion

In this section, we present a comprehensive discussion of the evaluation measures, experimental details, and the results obtained through the proposed methodology. We delve into the assessment criteria used to gauge the performance of our approach, provide insights into the experimental setup and configurations, and present the outcomes achieved during our evaluation process.

4.1. Evaluation Metrics

In the realm of machine learning and deep learning, evaluation metrics play a vital role in gauging model performance. These measures are fundamental in statistical research and are essential in assessing the effectiveness of our proposed model. In this study, we emphasized the following key assessment measures [40] to evaluate the efficacy of our approach. In Equations (9)–(12), *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

- **Accuracy:** Accuracy is a metric that assesses the overall correctness of the model's predictions. It calculates the proportion of correctly classified samples out of the total

samples. While accuracy is a crucial evaluation measure, it may not be sufficient in certain scenarios, such as imbalanced datasets or cases where different types of errors have varying consequences. In such situations, additional evaluation metrics may be necessary to provide a more comprehensive understanding of the model's performance and capabilities.

$$Accuracy = {}^{TP+TN}/_{TP+FP+TN+FN} \quad (9)$$

- **Precision:** Precision (P) is a metric that evaluates a model's capability to correctly identify positive samples among the predicted positive samples. It calculates the proportion of true positive predictions to the total number of positive predictions (which includes both true positives and false positives). Precision provides valuable insights into how accurately the model detects and classifies positive instances, making it an essential measure in many classification tasks.

$$P = {}^{TP}/_{TP+FP} \quad (10)$$

- **Recall:** Recall (R), alternatively termed sensitivity or the true positive rate, gauges the model's ability to accurately recognize positive samples within the total pool of actual positive samples. It is computed as the ratio of true positives to the sum of true positives and false negatives. Recall signifies the model's effectiveness in comprehensively capturing positive instances, rendering it a crucial assessment metric in classification tasks.

$$R = {}^{TP}/_{TP+FN} \quad (11)$$

- **F1 Score:** The *F1* score is computed as the harmonic mean of precision (P) and recall (R), providing a single statistic that balances the two metrics. This makes it particularly useful when dealing with imbalanced class distributions or scenarios where equal emphasis is placed on both types of errors. The *F1* score ranges from 0 to 1, with 1 representing the best possible performance of the model. By incorporating both precision and recall, the *F1* score offers a comprehensive evaluation of the model's overall effectiveness in classification tasks.

$$F1 = {}^{(2\times P\times R)}/_{(P+R)} \quad (12)$$

4.2. Results and Discussion

The ViT model was trained using various aspects of the prepared dataset. In the following sections, we present the classification report and learning graphs showcasing the model's performance and capabilities in addressing various challenges. The ViT model required a huge amount of data for training purposes and a 40K image-based dataset was effective for this study. Further, we considered 20% of the data for evaluation purposes. The dataset split ratio in terms of class balancing can be seen in Table 3.

Table 3. Train, validation, and test dataset splitting.

Dataset	Diffusion_Fake	GAN_Fake	Real	Stable&GAN_Fake	Total
Train	7203	7192	7203	7204	28,802
Validation	1800	1798	1800	1800	7198
Test	997	1010	997	996	4000

During the training process, the model's training loss initially started at 0.0457 and gradually decreased to 0.0020. Similarly, the validation loss began at 0.0235 and eventually reached 0.0079. The loss graphs, depicting the variations in loss to epochs, are illustrated in

Figure 4. These graphs provide valuable insights into the model's learning progress and ability to optimize the training process, ultimately improving performance and accuracy.

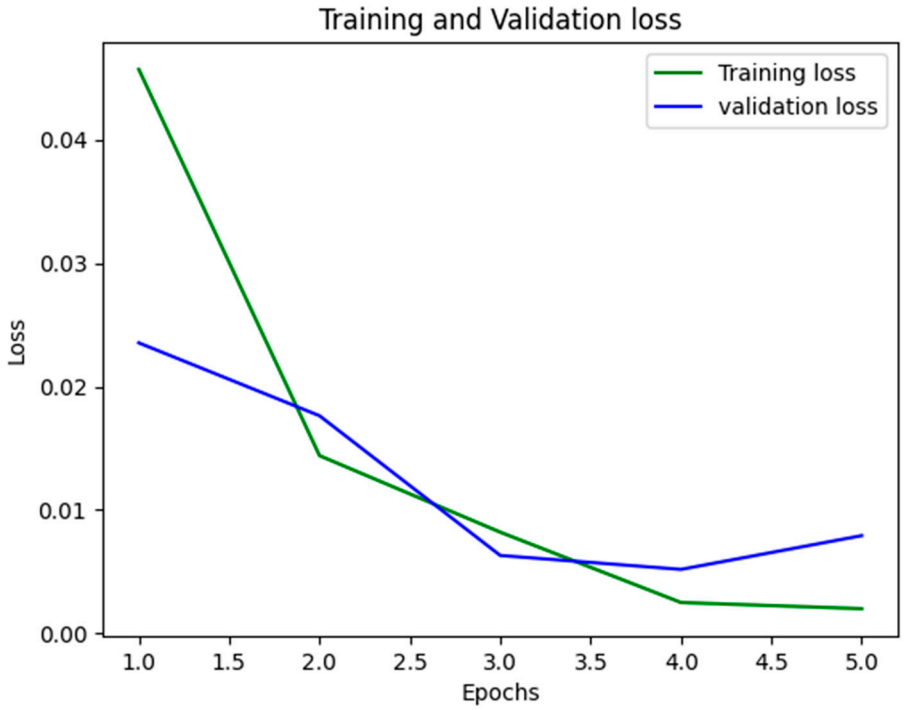


Figure 4. Training and validation loss of ViT model.

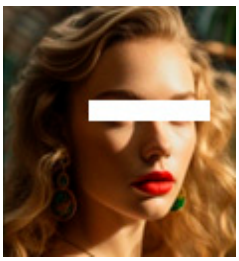
In our evaluation of the proposed model, we also employed class-wise precision, recall, and F1 score to assess its performance, as presented in Table 4. The support column indicates the number of samples available for each class in the testing dataset. For example, the Real class consists of 997 samples, and the Diffusion Fake class also comprises 997 samples for testing purposes. The total sum of the support column equals 4000, representing the total number of samples tested in our evaluation. By analyzing these class-wise metrics, we can understand the model's effectiveness in correctly classifying different classes and its overall performance across the entire dataset.

Table 4. ViT performance class-wise for multiclass deepfake identification.

Class Name	Precision	Recall	F1	Support
Diffusion_Fake	1.0000	1.0000	1.0000	997
GAN_Fake	1.0000	0.9960	0.9980	1010
Real	1.0000	1.0000	1.0000	997
Stable&GAN_Fake	0.9960	1.0000	0.9980	996

Table 5 showcases the actual and predicted labels with the ViT model. The table contains three columns: "Images", "Predicted", and "Actual". Each row corresponds to a different image. The "Predicted" column displays the labels that the ViT model assigned to the images after analyzing them, while the "Actual" column shows the true labels. Table 5 also demonstrates that the ViT model accurately predicted the labels for all the tested images, with its predictions matching the actual labels except for the last image. The excessive use of filters and a side pose could be the reasons for misclassification. This suggests that the ViT model is effective in classifying different image types based on the provided data. Furthermore, to test any image in the future, please follow the steps outlined in the https://github.com/Muhammad-Asad-Arshed/MultiClass_DeepFake.git (accessed on 25 November 2023) repository.

Table 5. Actual vs. predicted using ViT model.

Images	Predicted	Actual
	GAN_Fake	GAN_Fake
	Diffusion_Fake	Diffusion_Fake
	Stable&GAN_Fake	Stable&GAN_Fake
	Real	Real
	Diffusion_Fake	Real

1. Our proposed model weights will be downloaded.
2. Install the necessary libraries.
3. Extract the model weights that are in the RAR file and load the model.
4. Upload an image to Google Colab and set the path in the "img" variable.
5. Run the "Prediction" cell to get the class.

4.2.1. Comparison with CNN-Based Pretrained Architectures

To demonstrate its robustness and highlight the effectiveness of global feature extraction in deepfake identification over local feature extraction, our proposed model was

meticulously evaluated against established CNN-based models. This comparison serves to underscore the model's capability in capturing comprehensive patterns across the entire dataset, emphasizing its potential superiority in discerning deepfake content.

We achieved a training accuracy of 0.77, a train accuracy of 0.78, and a test accuracy of 0.77 with a fine-tuned ResNet-50 model [20]. The graphical representation of the learning graph can be seen in Figure 5.

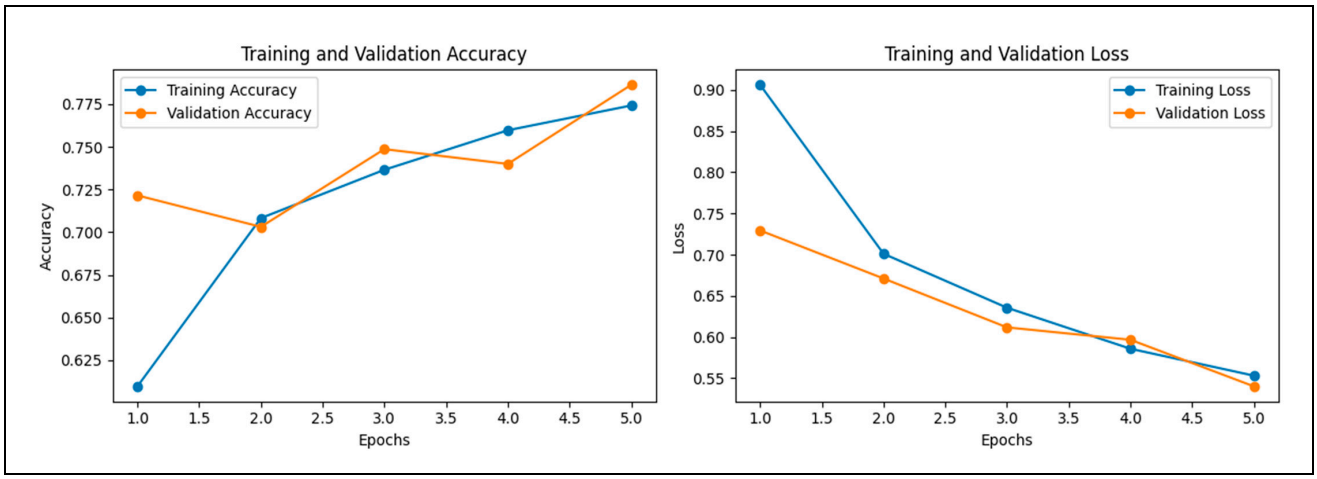


Figure 5. Training and validation graph of fine-tuned ResNet-50 model.

The fine-tuned VGG-16 model [19] has demonstrated noteworthy performance, achieving a training accuracy of 0.95 and a validation accuracy of 0.93 compared to the ResNet-50 model. The model's effectiveness extends to the test dataset, where it maintains a robust accuracy of 0.94. For a comprehensive visual representation of these results, see Table 6 and Figure 6, which illustrate the efficacy and reliability of the VGG-16 model.

Table 6. Proposed model comparison with local feature extraction-based pretrained models (based on two decimal place evaluation scores).

Model	Train Accuracy	Validation Accuracy	Accuracy	Precision	Recall	F1
ResNet-50	0.77	0.78	0.77	0.80	0.77	0.78
VGG-16	0.95	0.93	0.94	0.94	0.94	0.94
Proposed	0.99	0.99	0.99	0.99	0.99	0.99

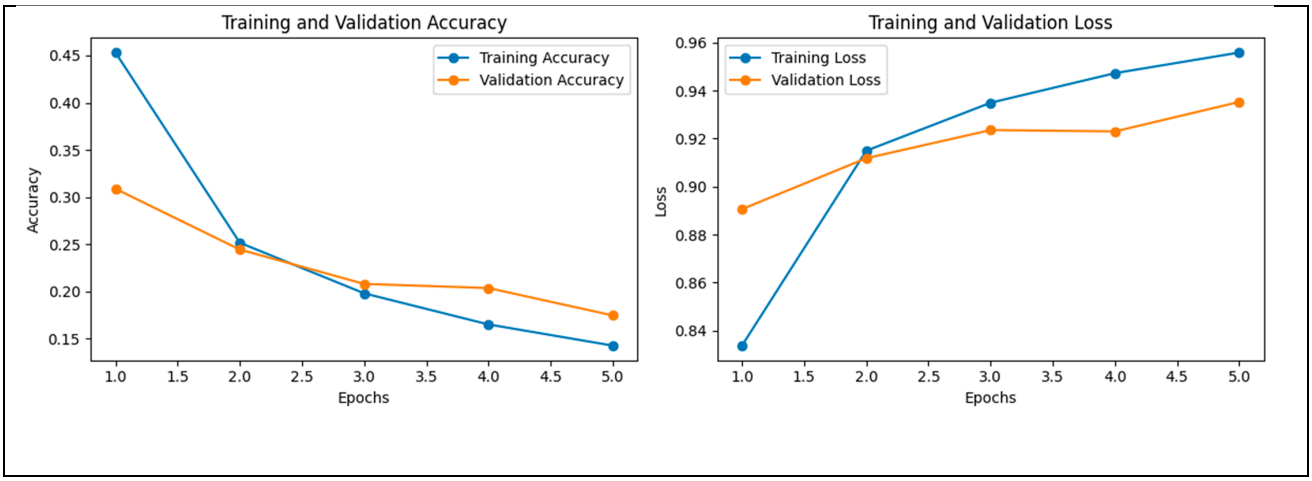


Figure 6. Training and validation graph of fine-tuned VGG-16 model.

4.2.2. Comparison with the Literature Contributions

It is important to acknowledge that a direct comparison with existing studies in the field of deepfake identification may not be feasible due to the unique nature of our research. Our study is foundational work specifically focusing on predicting deepfakes as a multiclass problem. In contrast, most existing studies are based on binary classification ([41–47]) distinguishing between real and fake images (see Table 7). As our approach tackles the complex task of multiclass deepfake identification, it introduces novel challenges and considerations that differentiate it from previous research. Therefore, caution should be exercised when drawing direct comparisons with binary-based studies, as these approaches' contexts and objectives differ significantly. Our research seeks to contribute to the field by exploring the capabilities and limitations of multiclass deepfake detection, paving the way for further advancements in this emerging study area.

Table 7. Analysis of the proposed study in contrast to existing state-of-the-art studies.

Authors	Methodology	Dataset	Classes	Accuracy (%)
(Gandhi et al., 2020) [41]	Pretrained ResNet Model	10,000 Images	2	94.75%
(Hu et al., 2021) [42]	Highlights of Corneal Specular	1000 Images	2	90.48%
(Yousaf et al., 2022) [43]	CNN Based on Two Stream	11,982 Images	2	90.65%
(Haung et al., 2022) [45]	Implicit Identity-Driven Framework employing Explicit Identity Contrast (EIC) and Implicit Identity Exploration (IIE) losses	10,000 videos DeepFakesofFF++(C23) and FaceShifter	2	67.99–88.21%
(Raza et al., 2022) [46]	Hybrid (VGG16 and CNN)	2041 Images	2	94%
(Arshed et al., 2023) [47]	Transformer	100,000 Images	2	99.5–100%
Proposed (Experiment 1)	ResNet-50	40,000 Images	4	77%
Proposed (Experiment 2)	VGG-16	40,000 Images	4	94%
Proposed (Experiment 3)	ViT	40,000 Images	4	99.90%

Additionally, this study holds significant importance by pioneering a multi-classification approach to deepfake detection, a previously unexplored aspect, thereby advancing the field's understanding and effectiveness in countering evolving deepfake challenges. The creation of a dedicated dataset for multiclass deepfake identification facilitates enhanced model training and accuracy. Introducing a fine-tuned ViT model that surpasses state-of-the-art techniques underscores the research's advancements. Moreover, this study establishes the proposed method's robustness and generalizability through extensive analysis, reinforcing its reliability for combating diverse deepfake scenarios and content types.

4.2.3. Implications

This study introduces a novel theoretical perspective by framing deepfake detection as a multiclass task, acknowledging the diversity in manipulation techniques like Stable Diffusion and StyleGAN2. The application of ViT for global feature extraction represents a theoretical advancement, expanding beyond traditional CNNs. Recognizing and addressing challenges posed by advanced techniques contributes to a nuanced understanding of deepfake intricacies. On a practical level, the proposed ViT-based method demonstrates exceptional accuracy (99.90%) on a multiclass-prepared dataset, highlighting its robustness in countering deepfake threats. The comparison with state-of-the-art CNN models provides a practical benchmark, emphasizing the ViT's superiority and contributing significantly to a more secure digital landscape.

5. Conclusions

Deepfakes have emerged as a prominent technique for disseminating misinformation and manipulating visual content. While not all deepfake creations are inherently malicious,

it is essential to identify and address such content, as some instances can pose significant threats to society. In this study, we focused on the critical task of multiclass deepfake identification and evaluated the effectiveness of the ViT in detecting deepfake images. The inherent global feature mapping and self-attention mechanisms of the ViT proved to be highly effective in discerning deepfake content. Through rigorous evaluation across various image manipulation and generation techniques, our approach achieved an exceptional accuracy of 99.90%. These results highlight the ViT's potential to combat deepfake content and promote trust and integrity in digital media. Our research endeavors will focus on expanding the scope of our current work by incorporating additional datasets specifically curated and released for deepfake research. This expansion is essential to enhance the diversity, accuracy, and overall robustness of our methods and findings and to address the ever-evolving challenges posed by deepfake technology. Our ongoing efforts strive to contribute to the advancement of deepfake detection and contribute to building a more secure and trustworthy digital landscape.

Author Contributions: Conceptualization, M.A.A.; methodology, M.A.A.; validation, M.T., C.D., S.M., M.I. and S.A.; supervision, S.M.; investigation, M.A.A., S.M., and S.A.; formal analysis, M.A.A., M.T. and M.I.; data curation, M.A.A.; writing—original draft preparation, M.A.A.; writing—review and editing, M.A.A. and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The prepared dataset will be provided on request.

Acknowledgments: The authors would like to thank all colleagues from Satya Wacana Christian University, Indonesia, and all involved in this research. This research is supported by the Vice-Rector of Research, Innovation and Entrepreneurship at Satya Wacana Christian University. We also acknowledge the use of ChatGPT (https://chat.openai.com/, (accessed on 25 November 2023)) for English correction across all manuscript sections.

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

Abbreviations

TP	True Positive
FP	False Positive
TN	True Negative
TP	True Positive
FN	False Negative
ViT	Vision Transformer
DARPA	Defense Advanced Research Projects Agency
GAN	Generative Adversarial Networks
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
AI	Artificial Intelligence
E	Embeddings
Pos	Positional

References

1. 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] [PubMed]
2. Zhang, C.; Zhang, C.; Zheng, S.; Zhang, M.; Qamar, M.; Bae, S.-H.; Kweon, I.S. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI. 2023. Available online: http://arxiv.org/abs/2303.13336 (accessed on 10 August 2023).
3. Wu, X.; Xu, K.; Hall, P. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Sci. Technol.* **2017**, *22*, 660–674. [CrossRef]
4. Wojewidka, J. The deepfake threat to face biometrics. *Biom. Technol. Today* **2020**, *2020*, 5–7. [CrossRef]

5. van der Sloot, B.; Wagensveld, Y. Deepfakes: Regulatory challenges for the synthetic society. *Comput. Law Secur. Rev.* **2022**, *46*, 105716. [CrossRef]
6. Gregory, S. Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism* **2021**, *23*, 708–729. [CrossRef]
7. AI Deepfake Videos: The Growing Concerns and Potential Harm—DevX. Available online: https://www.devx.com/news/ai-deepfake-videos-the-growing-concerns-and-potential-harm/ (accessed on 9 August 2023).
8. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (accessed on 11 July 2023).
9. How Deepfakes Deceptions are Affecting Businesses. Available online: https://shuftipro.com/blog/how-deepfakes-deceptions-are-affecting-businesses/ (accessed on 9 August 2023).
10. The Dangers of Deepfakes—A Cybersecurity Perspective—IPV Network. Available online: https://ipvnetwork.com/the-dangers-of-deepfakes-a-cybersecurity-perspective/ (accessed on 9 August 2023).
11. Diakopoulos, N.; Johnson, D. Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. *New Media Soc.* **2020**, *23*, 2072–2098. [CrossRef]
12. Nguyen, Q.V.H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.-V.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2022**, *223*, 103525. [CrossRef]
13. Tampubolon, M. Digital Face Forgery and the Role of Digital Forensics. *Int. J. Semiot. Law* **2023**, 1–15. [CrossRef]
14. Passos, L.A.; Jodas, D.; da Costa, K.A.P.; Júnior, L.A.S.; Rodrigues, D.; Del Ser, J.; Camacho, D.; Papa, J.P. A Review of Deep Learning-based Approaches for Deepfake Content Detection. *arXiv* **2022**, arXiv:2202.06095.
15. Media Forensics. Available online: https://www.darpa.mil/program/media-forensics (accessed on 9 August 2023).
16. Facebook, Microsoft Back Contest to Better Detect Deepfakes | WIRED. Available online: https://www.wired.com/story/facebook-microsoft-contest-better-detect-deepfakes/ (accessed on 9 August 2023).
17. Akhtar, Z.; Mouree, M.R.; Dasgupta, D. Utility of Deep Learning Features for Facial Attributes Manipulation Detection. In Proceedings of the 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI), Irvine, CA, USA, 21–23 September 2020; pp. 55–60.
18. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. 2016. Available online: https://arxiv.org/abs/1602.07360v4 (accessed on 12 July 2023).
19. 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; Available online: https://arxiv.org/abs/1409.1556v6 (accessed on 12 July 2023).
20. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
21. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
23. Agarwal, S.; Varshney, L.R. Limits of Deepfake Detection: A Robust Estimation Viewpoint. 2019. Available online: https://arxiv.org/abs/1905.03493v1 (accessed on 9 August 2023).
24. Akhtar, Z.; Dasgupta, D. A Comparative Evaluation of Local Feature Descriptors for DeepFakes Detection. Available online: https://ieeexplore.ieee.org/abstract/document/9033005/ (accessed on 11 July 2023).
25. 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), Gaziantep, Turkey, 5–7 October 2020; pp. 1–4.
26. Li, Y.; Chang, M.-C.; Lyu, S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
27. 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).
28. Patel, M.; Gupta, A.; Tanwar, S.; Obaidat, M.S. Trans-DF: A Transfer Learning-based end-to-end Deepfake Detector. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 796–801.
29. Ciftci, U.A.; Demir, I.; Yin, L. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September—1 October 2020; pp. 1–10.
30. Yang, J.; Xiao, S.; Li, A.; Lu, W.; Gao, X.; Li, Y. MSTA-Net: MSTA-Net: Forgery detection by generating manipulation trace based on multi-scale self-texture attention. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4854–4866. [CrossRef]
31. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. Available online: https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html?ref=https://githubhelp.com (accessed on 12 July 2023).

32. Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.-G.; Li, S.-N. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In Proceedings of the ICMR'22: International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022.
33. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A Survey of Visual Transformers. Available online: https://ieeexplore.ieee.org/abstract/document/10088164/ (accessed on 9 August 2023).
34. k Real and Fake Faces | Kaggle. Available online: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces (accessed on 12 July 2023).
35. thispersondoesnotexist.com (1024 × 1024). Available online: https://thispersondoesnotexist.com/ (accessed on 12 July 2023).
36. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 10674–10685.
37. Synthetic Faces High Quality (SFHQ) Part 4 | Kaggle. Available online: https://www.kaggle.com/datasets/selfishgene/synthetic-faces-high-quality-sfhq-part-4 (accessed on 2 August 2023).
38. 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. 2020. Available online: https://arxiv.org/abs/2010.11929v2 (accessed on 12 May 2023).
39. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
40. Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score | by Vaibhav Jayaswal | Towards Data Science. Available online: https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262 (accessed on 21 December 2023).
41. Gandhi, A.; Jain, S. Adversarial Perturbations Fool Deepfake Detectors. Available online: https://ieeexplore.ieee.org/abstract/document/9207034/ (accessed on 13 July 2023).
42. Hu, S.; Li, Y.; Lyu, S. Exposing GaN-generated faces using inconsistent corneal specular highlights. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2500–2504.
43. Yousaf, B.; Usama, M.; Sultani, W.; Mahmood, A.; Qadir, J. Fake visual content detection using two-stream convolutional neural networks. *Neural Comput. Appl.* **2022**, *34*, 7991–8004. [CrossRef]
44. Lyu, S. DeepFake Detection. In *Advances in Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 313–331. [CrossRef]
45. Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; Ye, D. Implicit Identity Driven Deepfake Face Swapping Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 4490–4499.
46. Raza, A.; Munir, K.; Almutairi, M. A Novel Deep Learning Approach for Deepfake Image Detection. *Appl. Sci.* **2022**, *12*, 9820. [CrossRef]
47. 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]