DeepFake Detection: Harnessing Neural Networks for Digital Image Integrity

Monika[1][0000−0002−6019−6493], Tannu Tiwari[2], Shaurya Singh[3], Arjun Singh[4], and Arya Yadav[5]

SCSET, Bennett University, Greater Noida, U.P., India[1,2,3,4,5]
monika@bennett.edu.in

Abstract. Deepfake technology, accelerated by advancements in deep learning, poses significant threats to digital media integrity. This study evaluates four key models: CNN, ResNet, Logistic Regression, and Vision Transformers (ViT) for their ability to detect deepfakes by analysing spatial features, fine-grained anomalies, and classification efficiency. This research compares the models based on accuracy, computational efficiency, and robustness, using datasets from FaceForensics++. The results highlight the superior accuracy of deep learning models but emphasise trade-offs in computational cost for real-time applications in areas such as social media monitoring and cybersecurity. This work advances digital forensics, aiming to combat misinformation and to uphold trust in AI-driven content.

Keywords: Deepfake detection · CNN · ResNet · Logistic Regression · Vision Transformers · FaceForensics++ · AI · digital forensics · real-time detection.

1 Introduction

With deepfake technology, unprecedented challenges for the authentication of digital media have arisen and penetrated profoundly into public confidence, privacy, and security (11). By leveraging sophisticated artificial intelligence (AI) techniques, particularly generative adversarial networks (GANs), deepfakes produce hyper-realistic media that can deceive even the most perceptive viewers (7). Originally developed for artistic and recreational purposes, such as virtual reality and filmmaking, deepfake technology has rapidly evolved into a powerful tool for malicious applications, including identity theft, defamation, and disinformation campaigns (12). These evolving threats highlight the urgent need for reliable detection methods to safeguard public trust and institutional integrity (4). Unlike traditional forms of digital manipulation, deepfakes manipulate content at the pixel level, making it rather difficult to detect (13). Some early detection techniques sought to identify visual artifacts commonly used in first-generation deepfakes, such as unnatural eye movements, inconsistent lighting, or distorted facial features (2). However, recent upgrades in algorithms like StyleGAN and Face2Face made these imperfections harder to detect (19). Hence, it is

now required to employ advanced detection methods, which can identify subtle mismatches in spatial as well as temporal dimensions. The investigation on machine learning-based deepfake detection emphasizes using the method of CNN to look into spatial analysis (9). Small problems that people might not notice otherwise, like slightly misaligned faces and irregular heartbeats that can be seen in tiny expressions on a person's face, were tested (10). Audio-visual synchrony in a multi-modal detection system can show lip-sync and speech problems that a person can't see or doesn't notice with their own eyes (15).

A key focus area of this research is the designing of a scalable and robust deepfake detection framework employing publicly available datasets such as FaceForensics++ and the Deepfake Detection Challenge.(8) These datasets provide a variety of high-quality samples for training, enabling researchers to evaluate the efficacy of the proposed models in different types of deepfakes (18). Finally, the paper discusses the pressing requirement for real-time detection capabilities to ensure that the suggested framework can be implemented in high-risk areas such as digital content verification, social media supervision, and national security activities. In addition to being a technical advance, this study adds to the field of digital forensics because it shows how important ethical AI is for lowering the risks that deepfakes pose to society. The development of strong detection systems aims to combat the dissemination of misinformation, uphold democratic processes, and safeguard privacy.

The rest of the paper is structured as follows: Section 2 addresses the efforts undertaken in the realm of detecting false news, while the proposed approach is delineated in Section 3, followed by the analysis of results in Section 4 and closing observations in Section 5.

2 Literature Review

Recent research has focused on developing deep neural network-based systems for detecting deepfake images and videos. Various architectures have been explored, including combinations of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks.

Agnihotri (1) investigated CNN-LSTM models for deepfake detection and reported accuracies ranging from 86% to 98%. The study also explored commonly used CNN architectures such as EfficientNetB4, InceptionV3, Inception-ResNetV2, and ResNet50. Additionally, publicly available datasets such as Flickr Faces High Quality (FFHQ), Celeb-DF, and Face Forensic++ were used for training and testing . Rani B R et al. (6) further examined CNN architectures, including EfficientNetB4 and ResNet50, for detecting deepfakes. The study utilized datasets like FFHQ and Celeb-DF to evaluate model performance. Saini et al. (16) looked into CNN-LSTM-based models for deepfake detection, focusing on the steps needed to find the fakes, such as frame extraction, face detection, and feature vector generation. The study reported promising detection accuracies similar to previous works.

Even though there have been improvements, Badale et al. (5) pointed out how hard it is to tell the difference between increasingly complex deepfakes and real media, showing the need for ongoing improvements in detection methods. According to Altaei and Mohammed Sahib Mahdi (3), they suggest using convolutional neural networks (CNNs) and principal component analysis (PCA) along with deep learning to find deepfake images. He tests this method with and without PCA and finds that CNN-based methods can accurately find deepfake images with larger datasets and without preprocessing. Hasin Shahed Shad et al. (17) implemented and compared the performance of several CNN models like ResNet50, DenseNet201, DenseNet169, VGG19, VGG16, and DenseNet121 for detecting deepfake images from real ones, with the VGGFace model performing the best with 99% accuracy. S. Pashine et al. (14) compare the performance of four deep learning models—MesoNet, ResNet-50, VGG-19, and Xception Net—to find the best one for different situations, like finding things in real time on social media or with a high level of accuracy for news organizations.

The research highlights the dominance of convolutional neural networks in the detection of deepfake content. Techniques like EfficientNet and ResNet have been widely employed to extract subtle artifacts inherent in manipulated images and videos. The ability to generalize to unfamiliar deepfake algorithms is a continual concern. Numerous detection systems excel on recognized datasets but encounter difficulties with new techniques of content creation, underscoring the necessity for adaptive and resilient models.

3 The Proposed Methodology

This paper used four different models: CNN, ResNet, logistic regression, and Vision Transformer for deepfake detection. Since FaceForensics++ includes varying deepfake samples, this is expected to validate the various strengths and weaknesses among the age-old traditional ML methods along with new advancements in the DL field of deepfakes. The methodology includes data preprocessing, model development, training and fine-tuning, and performance evaluation. Detailed descriptions of each stage are provided below.

3.1 Dataset Description

To test and train the deepfake image detection models, we utilized some of the most well-known datasets that boast a diverse variety of manipulations used in creating such images. We used datasets like FaceForensics++, Celeb-DF, and Deepfake Detection Challenge (DFDC). This dataset includes both real pictures and pictures that were made by computers using techniques like face swapping with GAN-based models, re-enacting expressions, and full-body synthesis. These varieties facilitate all-round testing for detection models regarding different kinds of manipulations. Each dataset poses unique challenges: FaceForensics++: It contains high-resolution face-swapped images that challenge models to identify pixel-level artifacts and facial inconsistencies as shown in figure 1. Celeb-DF: A

Fig. 1: FaceForensics++ Image Data set Sample

high-quality, realistic image of celebrities, challenging models with subtle manipulation cues. The DFDC dataset is extremely large, with a rich variety of manipulation techniques as well as compression levels to test model robustness within real-world scenarios.

3.2 Dataset Preparation

For the purpose of enabling reliable detection of deepfake images, a structured approach to the production of the dataset is utilised. This strategy ensures that the data retains both diversity and consistency.

1. Frame Extraction: Videos are converted into individual frames to provide static images for analysis. For consistency, only a subset of frames (e.g., every 10th frame) is used to balance dataset size and variability.
2. Labeling: Each frame is labeled as "real" or "fake" based on its origin.
3. Data Augmentation: Techniques like rotation, flipping, zooming, and color jittering are applied to increase diversity and prevent overfitting.
4. Normalization: Images were normalized to have the same pixel distribution so that the model converged.
5. Image Sampling: In the case of video frame datasets, representative frames were taken so that the key artifacts visible in still images were focused upon.

6. Splitting: The dataset is split into three parts: one is Training set (80% of data), Validation set (10% of data) and test set (10% of data), as shown in Figure 2

3.3 Detection Models

Convolutional Neural Networks : CNN are a class of deep learning models adept at extracting hierarchical spatial features from images. For this study, a custom CNN is designed with three convolutional layers (each followed by batch normalisation, ReLU activation, and max-pooling) and two fully connected layers as shown in Figure 3. Dropout is applied to prevent overfitting.

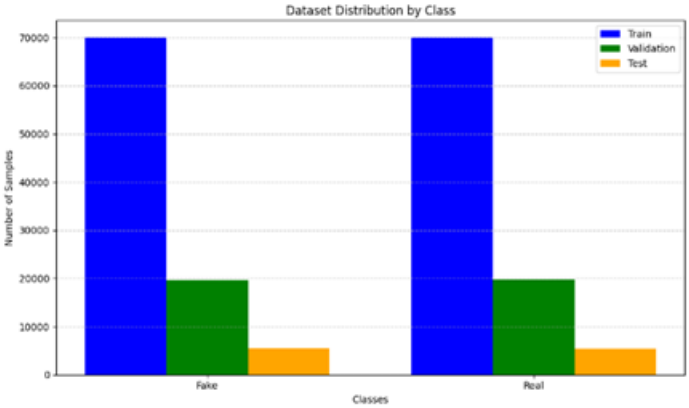


Fig. 2: Splitting of Dataset

The preprocessed images will be the input from FaceForensics++. and each image is categorised as either "real" or "fake". The CNN model is consists of layers that extract hierarchical features from images. Let X represent the input image and W the weights of the convolutional filters. The output Y of each convolutional layer can be expressed as:

$$Y = f(W * X + b) \tag{1}$$

where $*$ denotes the convolution operation, b is the bias term, and f is the activation function. Here ReLU function is used. The CNN is trained using the Backpropagation (BP) algorithm, where, in the forward propagation phase, the output of each layer is produced. Then the error is calculated using cross entropy loss.

$$L = -\sum (y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{2}$$

In the backpropagation phase, weights were updated using gradient descent.

Residual Neural Network : ResNet is a deep learning architecture designed to overcome the vanishing gradient problem, enabling efficient training of very deep networks. It is highly effective for binary classification tasks. ResNet assumes that deeper networks can extract more complex features through multiple layers. Instead of learning a direct mapping, ResNet learns residual functions to address the problem of vanishing gradients. ResNet introduces a residual block where the input XX bypasses the convolutional layers through a shortcut connection, resulting in the output:

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

$f(X)$ is the transformation applied by convolutional layers and X is the identity shortcut connection. For each residual block, the transformed output becomes:

$$Y = f(W * X + b) + X \tag{4}$$

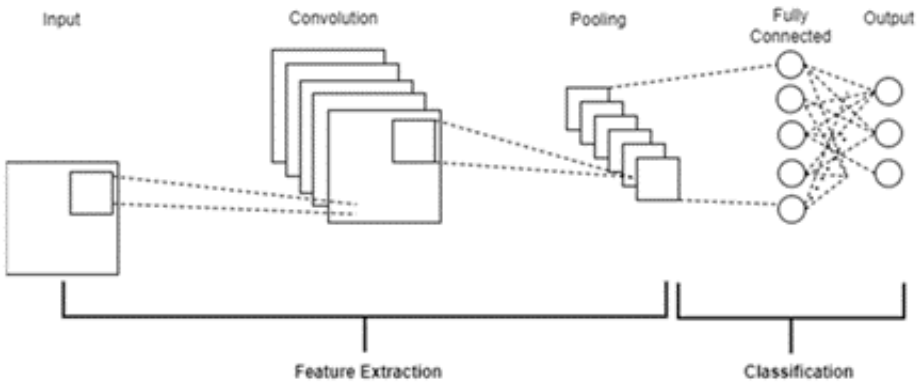


Fig. 3: Deepfake Image Detection & Classification using Conv2D Neural Networks

where W and b are the weights and biases of the convolutional layers. Here, ReLU (Rectified Linear Unit) is used as an activation function. Before classification, the features from the last convolutional layer are put through global average pooling (GAP) to make the space smaller. Finally, a fully connected layer followed by a SoftMax or sigmoid activation is applied to predict the class probabilities. The major are discussed as follows:

1. Input: Pre-processed images from the FaceForensics++ dataset.
2. Architecture: A ResNet model comprises stacked residual blocks, incorporating convolutional layers, batch normalization, and ReLU activation functions. Skip connections facilitate the learning of residual mappings, thereby enhancing training stability and accuracy.
3. Optimization: Cross-Entropy Loss is used for model training, and fully connected layers are subjected to Dropout in order to minimize overfitting.
4. Hyperparameters: To achieve the best results on the binary classification job, the learning rate, batch size, and dropout rate are adjusted.
5. Output: A binary classification result denoting the authenticity of a picture as either real or counterfeit.

Logistic Regression : Logistic regression provides a simple yet effective baseline for binary classification. Feature vectors extracted from pretrained CNNs. Logistic Regression uses the sigmoid function to predict probabilities for the two classes (real and fake) as shown in the equation, where w is the weight vector, X the feature vector, and b the bias term.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(w \cdot X + b)}} \tag{5}$$

The algorithm minimizes log loss during training. Logistic Regression uses the binary cross-entropy Loss, identical to CNN's loss function.

Vision Transformer (ViT): Vision Transformers (ViTs) represent advanced models that employ self-attention mechanisms to effectively capture long-range dependencies within images. A ViT pretrained on ImageNet is fine-tuned on the

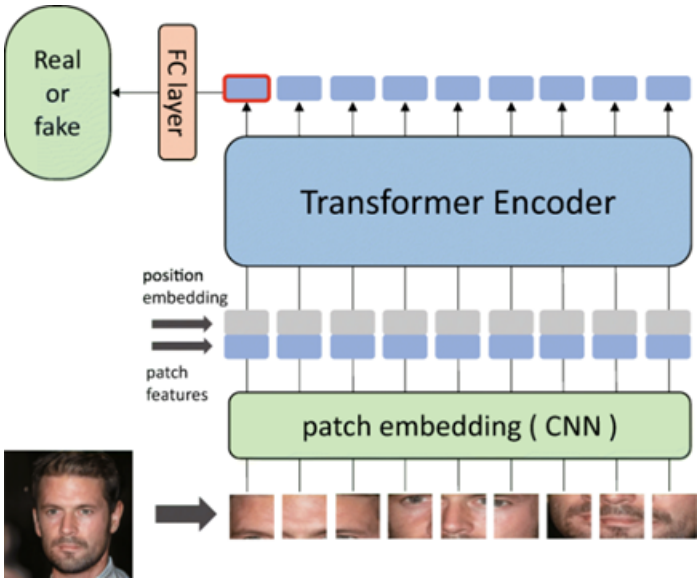


Fig. 4: DeepFake detection algorithm based on improved vision transformer

FaceForensics++ dataset. Images are split into patches (e.g., 16×16, 16×16), which are linearly embedded and passed through multi-head attention layers. Unlike CNNs, ViTs capture global context more effectively, potentially identifying subtle manipulations missed by CNNs. Input images are divided into patches processed independently, and each patch Pi is linearly transformed using an equation, where W_p is the patch embedding matrix. It captures long-range dependencies.

$$Pi = XW_p + b_p \tag{6}$$

Every 10th frame is used to balance dataset size and variability. In a self-attention mechanism, attention scores is calculated using an equation. Techniques like rotation, flipping, zooming, and colour jittering are applied to increase diversity and prevent overfitting. The final transformer layer output is passed to an MLP head for classification. The functioning of the improved vision transformer is shown in figure 4. Unlike CNNs, ViTs capture global context more effectively, potentially identifying subtle manipulations missed by CNNs.

The effectiveness of the detection models was measured using an extensive set of evaluation metrics:

Accuracy: It calculated the ratio of correctly classified instances, providing a baseline comparison across models and datasets.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{TrueNegative} + \text{FalsePositive} + \text{FalseNegative}}$$

Precision, Recall, and F1-Score: These metrics assessed the ability of the models to reduce false positives and false negatives to ensure reliable identification of manipulated images.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This methodology employs a comprehensive approach to the identification of deepfake images by using diversity in datasets, advanced models of detection, ensemble techniques, and robust evaluation metrics. The framework aims to address the evolving challenges that arise from deepfake technology.

4 Results And Analysis

The models were trained using the FaceForensics++ dataset. To enhance model robustness, augmentation techniques such as rotation, flipping, and scaling were employed to increase the diversity of training data. An independent validation generator was employed to guarantee impartial assessment throughout the training process. Table 1 summarises the training metrics for all models. To evaluate

Table 1: Training metrics for all models

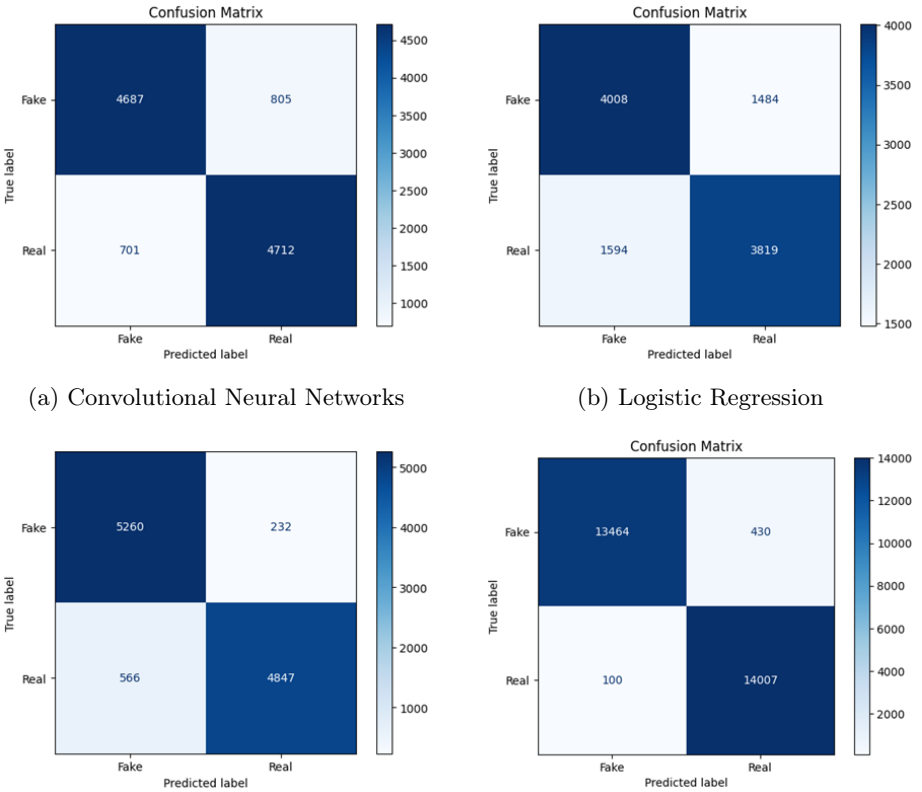
Model	**Training Loss**	**Validation Loss**	**Validation Accuracy (%)**
CNN	0.2259	0.2949	87.16
Logistic Regression	7.0822	6.9872	73.91
ViT	0.0294	0.0528	98.11
ResNet	0.0552	0.0643	95.33

generalization, the models were tested on an unseen dataset. Table 2 shows the test accuracy and loss values for each model.

The analysis results clearly demonstrate the effectiveness of neural network based models in deepfake image detection, particularly transformer-based architectures like ViT, which significantly outperformed traditional models. The effectiveness of ViT is due to its capacity to capture long-range dependencies

Table 2: Accuracy and Loss Values

Model	**Test Loss**	**Test Accuracy (%)**
CNN	0.3242	86.19
Logistic Regression	7.1421	71.77
ViT	0.0341	98.81
ResNet	0.2318	89.55



(a) Convolutional Neural Networks

(b) Logistic Regression

(c) Residual Neural Network

(d) Vision Transformer

Fig. 5: Confusion Matrix of Models

and complex features in images. ResNet also exhibited strong performance due to its deep hierarchical structure and residual learning approach.

However, the moderate performance of CNN suggests that while convolutional networks are effective, they may require additional fine-tuning or hybrid approaches (e.g., CNN-LSTM) to improve robustness. The poor results from logistic regression reaffirm the necessity of deep learning techniques for handling high-dimensional, complex visual data.

5 Conclusion

This research assessed the efficacy of deep neural network models for the detection of deepfake images utilizing the FaceForensics++ dataset. The results demonstrate that deep learning architectures, particularly Vision Transformer (ViT) and ResNet, significantly outperform traditional machine learning models in detecting deepfake content. ViT achieved the highest test accuracy of 98.81%, followed by ResNet with 89.55%, highlighting the superiority of transformer-based and deep CNN models in handling complex visual data.

Conversely, traditional models such as Logistic Regression performed poorly, reaffirming the necessity of advanced deep learning techniques for robust deepfake detection. While CNN models showed moderate performance, further improvements through hybrid architectures or additional fine-tuning could enhance their effectiveness. The results highlight the increasing difficulties presented by deepfake technology and the urgent requirement for more advanced and flexible detection techniques. Future research should explore hybrid deep learning approaches, adversarial training, and real-world deployment strategies to improve detection accuracy and resilience against increasingly sophisticated deepfake generation techniques.

Bibliography

[1] Agnihotri, A.: DeepFake Detection using Deep Neural Networks. Ph.D. thesis, Dublin, National College of Ireland (2021)

[2] Alanazi, S., Asif, S.: Exploring deepfake technology: creation, consequences and countermeasures. Human-Intelligent Systems Integration pp. 1–12 (2024)

[3] Altaei, M.S.M., et al.: Detection of deep fake in face images using deep learning. Wasit Journal of Computer and Mathematics Science **1**(4), 60–71 (2022)

[4] Awodiji, T.O., Owoyemi, J.: Advanced detection and mitigation techniques for deepfake video: Leveraging ai to safeguard visual media integrity in cybersecurity

[5] Badale, A., Castelino, L., Darekar, C., Gomes, J.: Deep fake detection using neural networks. In: 15th IEEE international conference on advanced video and signal based surveillance (AVSS). vol. 2 (2018)

[6] BR, S.R., Pareek, P.K., Bharathi, S., Geetha, G.: Deepfake video detection system using deep neural networks. In: 2023 IEEE international conference on integrated circuits and communication systems (ICICACS). pp. 1–6. IEEE (2023)

[7] George, A.S., George, A.H.: Deepfakes: the evolution of hyper realistic media manipulation. Partners Universal Innovative Research Publication **1**(2), 58–74 (2023)

[8] Godase, P., Pawar, S., Dhengale, R., Gurav, S., Kapare, A., Patel, R.: Deepfake detection using machine learning techniques: A scalable solution for media integrity (2024)

[9] Guarnera, L., Giudice, O., Battiato, S.: Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 666–667 (2020)

[10] Hohman, M.H., Kim, S.W., Heller, E.S., Frigerio, A., Heaton, J.T., Hadlock, T.A.: Determining the threshold for asymmetry detection in facial expressions. The Laryngoscope **124**(4), 860–865 (2014)

[11] Karnouskos, S.: Artificial intelligence in digital media: The era of deepfakes. IEEE Transactions on Technology and Society **1**(3), 138–147 (2020)

[12] Luz, H., Hery, A., Joseph, O., Femi, O.: Evolution of deepfakes: From recreation to malicious manipulation (2024)

[13] Naitali, A., Ridouani, M., Salahdine, F., Kaabouch, N.: Deepfake attacks: Generation, detection, datasets, challenges, and research directions. Computers **12**(10), 216 (2023)

[14] Pashine, S., Mandiya, S., Gupta, P., Sheikh, R.: Deep fake detection: survey of facial manipulation detection solutions. arXiv preprint arXiv:2106.12605 (2021)

[15] Rosemann, S., Wefel, I.M., Elis, V., Fahle, M.: Audio–visual interaction in visual motion detection: synchrony versus asynchrony. Journal of optometry **10**(4), 242–251 (2017)
[16] Saini, M.L., Patnaik, A., Sati, D.C., Kumar, R., et al.: Deepfake detection system using deep neural networks. In: 2024 2nd International Conference on Computer, Communication and Control (IC4). pp. 1–5. IEEE (2024)
[17] Shad, H.S., Rizvee, M.M., Roza, N.T., Hoq, S.A., Monirujjaman Khan, M., Singh, A., Zaguia, A., Bourouis, S.: [retracted] comparative analysis of deepfake image detection method using convolutional neural network. Computational intelligence and neuroscience **2021**(1), 3111676 (2021)
[18] Song, W., Yan, Z., Lin, Y., Yao, T., Chen, C., Chen, S., Zhao, Y., Ding, S., Li, B.: A quality-centric framework for generic deepfake detection. arXiv preprint arXiv:2411.05335 (2024)
[19] Taeb, M., Chi, H.: Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy **2**(1), 89–106 (2022)