Improving Vision Transformer for Deepfake Detection

Orvis L. Siagian, Reinhard Ebenhaizer, Pandu Wicaksono *, and Zahra N. Izdihar

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: orvis.siagian@binus.ac.id (O.L.S.); reinhard.ebenhaizer@binus.ac.id (R.E.);
pandu.wicaksono005@binus.ac.id (P.W.); zahra.izdihar@binus.ac.id (Z.N.I.)
*Corresponding author

***Abstract*—Machine learning is rapidly advancing across various fields and accelerating a paradigm shift in image and video manipulation. Deepfakes represent one of the challenges emerging from this development. Deepfakes are synthetically manipulated media using deep learning algorithms. Criminals have abused deepfakes as a weapon to spread false information. The distribution of deepfake videos or images may lead to some significant public risks, such as misleading information, privacy violation, and misuse in political and social realms. Therefore, the development of a counter for those threats is needed, namely a reliable deepfake detection method. One of the promising methods in the deepfake detection cases is the Vision Transformer (ViT). ViT is a deep learning architecture that uses self-attention mechanisms to understand complex relationships between images. Despite its potential, ViT needs a substantial amount of computational costs and a large dataset, which pose challenges for development. In this research, we present a rigorous evaluation of the ViT model with the use of the balanced FaceForensics++ dataset and 5-fold cross-validation strategy to ensure a more reliable result. The result shows an average accuracy of 85.39%, meaning that the model achieves a robust and stable performance. The model also showed an excellent balance between precision score (85.40%) and recall score (85.39%), which suggests to us that it is a reliable method in detecting deepfakes without significant bias. These findings indicate that a properly trained ViT, particularly with a balanced dataset, can serve as an effective and powerful tool to combat the threats posed by deepfakes.**

***Keywords*—vision transformer, deep learning, deepfake, machine learning, video manipulation**

I. Introduction

Deepfakes are photos or movies that have been altered using deep learning algorithms [1]. Deepfake content is typically manipulated sequences from videos of humans performing an action [2]. The rise of those threats raises concern in society, especially on the internet. By abusing this technology, perpetrators have caused various negative consequences, including misinformation, identity theft, pornography, and extortion [3]. The alteration involves switching the face of a person in the video with that of another individual. The alteration involves switching the face of a person in the video with that of another individual.

An example is the video featuring movie snippets from well-known films where the actors' faces were digitally manipulated to swap them onto the faces of others [4]. These movies are not intended for enjoyment, as users' faces may be captured unlawfully. This study indicates that the impact of deepfake video distribution is harmful [5]. One of the examples is political campaign videos. In the context of a black campaign, deepfakes are altered videos that show the speaker expressing controversial remarks. The alteration usually involves a swap between the speaker's face and a political candidate [6].

Given the situations, the significance of tools capable of detecting fake content in images and videos is essential [6]. We expect that deepfake detection algorithms will be able to be a valuable tool in protecting individuals against the exploitation of deepfake technology [1]. Numerous research studies have recently focused on advancing deepfake detection techniques for videos and images [1]. Deepfake detection techniques' objective is to categorize video or image content as original or manipulated [6]. The deepfake detection algorithms are built to look for characteristics in image or video content to recognize the difference between original and manipulated content [6]. The Vision Transformer (ViT) is one of the promising models to detect deepfake images or videos. ViT is an adaptation of a successful transformer architecture in Natural Language Processing (NLP). Unlike conventional Convolutional Neural Network (CNN), ViT employs a self-attention mechanism to understand the global interrelations between different segments of an image. The process involves dividing the image into fixed-size patches, which then are linearly embedded and processed by a transformer encoder. This encoder uses a multi-head self-attention mechanism to create a comprehensive functional representation of the image [7].

ViT has demonstrated superiority in comprehending global relationships among visual components. Recent studies have demonstrated the efficacy of ViT models in identifying counterfeit photos and movies, as evidenced by

Manuscript received August 28, 2025; revised September 18, 2025; accepted October 30, 2025; published February 27, 2026.

doi: 10.18178/joig.14.1.76-83

multiple investigations. Ghita *et al*. [8] conducted a study utilizing the ViT model to analyze 40,000 photos, achieving an accuracy of 89.9125%, demonstrating its potential while also highlighting areas for enhancement. The Vision Transformer's primary advantage is its superior ability to comprehend global relationships among visual components. Its strength is validated by recent studies, one of them is by Ghita *et al*. [8] which successfully utilizes the ViT model to analyse 40,000 photos, achieving a notable accuracy of 89.9125%. The study demonstrates its effectiveness in identifying fake content. Despite these promising results, this model presents significant challenges. The architecture is notoriously "data-hungry", requiring extensive datasets for effective training. Furthermore, it has shown limitations in capturing local-level information and suffers from high computational complexity, especially as input sizes increase [7, 8].

To address the challenges of deepfake detection, this paper provides a systematic and rigorous evaluation of the Vision Transformer (ViT), aiming to establish a dependable assessment of its efficacy as a countermeasure.

II. Literature Review

A. Deepfake

Deepfake is a combination of the word deep, which refers to Artificial Intelligence (AI) deep learning technology, and the word fake, which indicates that the content is not real [9]. The term Deepfake has been used since 2017 when Reddit moderators created a subreddit called "deepfakes" and posted videos that used face swap technology to insert celebrity likenesses into pornographic videos [10]. Recent research combines vision transformer with feature extraction from Convolutional Neural Networks (CNNs) such as DenseNet, creating an ensemble approach that improves deepfake detection capabilities in both images and videos. The collaboration of the DenseNet and Cross-ViT architecture shows superior results in distinguishing manipulated images with precision [11].

B. Vision Transformer

Vision Transformer (ViT) is a development of the Transformer NLP architecture for computer vision tasks. Vision Transformer adapts a self-attention mechanism which allows this model to understand global interactions between image parts without using kernel convolution as in Convolutional Neural Network. Various variations of Vision Transformer (ViT) were developed such as Data efficient image Transformer (DeiT) [12] and Convolutional Vision Transformer (ConViT) [13], namely locality-based models incorporating the local properties of CNN which aim to increase data efficiency. Then, Pyramid Vision Transformer (PVT) and Swin Transformer are hierarchical-based models that aim to reduce feature size gradually to increase computational efficiency. Apart from that, T2R-ViT and DeepViT are feature diversification-based models that focus on diversifying feature representations to improve performance [1].

C. Related Works

1) Deepfake detection based on visual features

This method is based on observable face features, like head posture, eye blink, and differences in the structure of facial organs. The studies employing this method were conducted in 2018 with eye blink as the primary characteristic. The method's underlying hypothesis is that the deepfake's blink pattern differs from the original video [6].

In another study, inconsistent head positions were used to identify deepfakes [1]. The methodology examines inconsistencies between the position of the face and non-facial body parts, such as the neck and shoulders [14]. Alternative methods are used to extract the "visual artifacts" produced by deepfakes. This term is used to describe the visual characteristics of faces created by deep faces. The characteristics seen from visual artifacts are the differences in color between the left and right eyes, the disproportionate shadows, geometric inaccuracies, and the details of light reflections. Due to the limited resources for creating deep fakes, we extract visual artefacts from their imperfections [6, 15, 16].

To identify deepfakes, techniques such as color and geometric extraction are employed to extract visual artifacts from specific facial features, including the nose, lips, teeth, eyes, and eyebrows [2]. However, as deepfakes become more sophisticated, it becomes difficult to identify these visual characteristics. As a result, visual feature-based detection techniques lose their effectiveness [17].

2) Deepfake detection based on local features

Another technique that can be used is feature extraction from each pixel in the image using pixel-based segmentation. This method is more reliable when compared to visual feature extraction on the face [8].

Zhou *et al*. [18] tried to propose a two-stream neural network that combines features from image stenography analysis, which captures residual local noise and camera characteristics, with features from a convolutional neural network, aiming to find manipulated areas in facial images. This study sets the foundation for deepfake detection techniques based on local and deep features. An additional strategy used is the Photo Response Non-Uniformity (PRNU) analysis with cross-correlation operations. However, this study only covered ten videos.

Additional feature extraction methods that can be applied are Binary Image Statistical Feature (BSIF), Oriented Gradient Histogram Pyramid (PHOG), Local Phase Quantization (LPQ), Local Binary Pattern (LBP), Accelerated Robust Feature (SURF), Binary Gabor Pattern (BGP), and Image Quality Metric (IQM). IQM is the most effective technique for identifying deepfakes, according to a study that compared it with Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) and discovered that it was the most useful feature [2].

In some video data, it has been discovered that local feature-based detection methods are quite effective at spotting deepfakes [2]. But as deepfake algorithms improve, modified information is harder to spot as deepfakes and more realistic. More sophisticated features are required to differentiate between original photos and films [19].

3) Detection of deep fakes based on deep features

The process of deep feature extraction mirrors that of local feature extraction. Due to the utilization of multiple layers in deep feature extraction, it can capture more intricate characteristics compared to conventional feature extraction methods. Sharafudeen *et al*. [20] investigated DenseNet, InceptionNet, and XceptionNet for the purpose of detecting deepfake images.

Several other studies have also compared deepfake detection using convolutional neural networks. For example, Ashok and Joy used the XceptionNet model on the FaceForecensics++ dataset, reporting over 95% accuracy on uncompressed and high-quality videos, and over 80% accuracy even on highly compressed videos [21]. These results set a strong performance benchmark for CNN-based approaches.

A deep autoencoder serves as an advanced model in the realm of deep learning, specifically designed for the detection of deepfakes. The autoencoder demonstrates the capability to identify and reconstruct elements of deepfake content [14]. One branch of the autoencoder successfully reconstructed the identified portion of the deepfake [22].

4) Deepfake detection based on temporal features

Recent studies have underscored the significance of temporal characteristics in deepfake detection, utilizing architectures like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and 3D Convolutional Neural Networks (3D CNN) to identify manipulation patterns over video frames. For instance, a solution put up in 2018 effectively employs a temporal-aware system that integrates a CNN with a convolutional LSTM network. This algorithm is meant to find strange things in deepfake movies that happen over time, like strange patterns of blinking and artifacts on the edges of the face. Combining ViT with modules that can handle temporal data could greatly increase the accuracy and generalization of deepfake video recognition, which is becoming more advanced [23].

III. Methodology

This section explains the methodology used to refine the Vision Transformer (ViT) model in deepfake video detection. In this section, there are six stages, including dataset selection, description of the pre-trained model, data pre-processing steps, dataset division, model refinement process, and evaluation metrics used.

A. Datasets

The dataset used for this experiment is a curated subset from the FaceForensics++ dataset, sourced from Kaggle (“Faceforensics-1000”) [24]. To address the class imbalance issues noted in previous experiments, this study utilized a balanced dataset consisting of 2000 videos in total. The collection is evenly split between 1000 authentic (‘real’) videos and 1000 manipulated (‘fake’) videos. This balanced approach is crucial for training a generalizable model and preventing bias towards any single class.

The complete FaceForensics++ dataset is large, and it takes a lot of computing power to process it all. As a result, we used the “Faceforensics-1000” curated subset for this investigation. This subset is a manageable but representative sample of the data. We chose to make a perfectly balanced dataset by picking an equal number of real and fake videos instead of using common methods for dealing with class imbalance like oversampling or the Synthetic Minority Over-sampling Technique (SMOTE). This choice was chosen to make sure that the model was only trained on real-world data with high integrity. This way, there would be no artifacts from synthetic data generation, and the model’s performance could be evaluated more accurately.

This dataset contains two categories of videos:

- Original (real) videos: Authentic video sequences that have not been tampered with any modifications.
- Manipulated (fake) videos: Video sequences that have been undergone deepfake alterations.

Example frames of the manipulated (‘fake’) and real (‘real’) videos from this dataset are shown in Figs. 1–4.

Frames from: 07_21__walking_down_street_outside_angry__K7KXUHMU.mp4
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5

Fig. 1. Manipulated video sample 1 frames.

Frames from: 16_28__outside_talking_pan_laughing__OANAQYD5.mp4
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5

Fig. 2. Manipulated video sample 2 frames.

Fig. 3. Original video sample 1 frames.

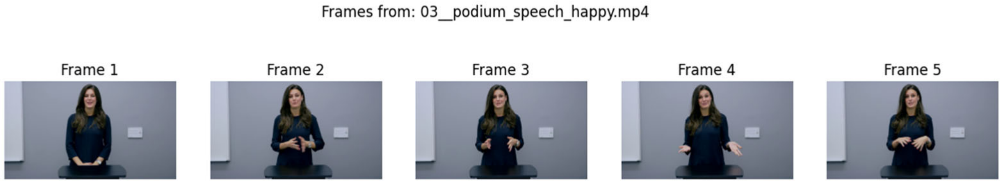

Fig. 4. Original video sample 2 frames.

B. *Pre-Trained Model*

This model is an implementation of Vision Transformer that processes images by dividing them into patches of 16×16 pixels and applying a self-attention mechanism [25].

The model has undergone pre-training on the ImageNet-21k large-scale dataset [26] and further fine-tuning on the ImageNet-1k dataset (ILSVRC2012) [27], both utilising an input resolution of 224×224 pixels.

The configuration used of the model includes:

- Patch size: 16×16.
- Input image resolution: 224×224 pixels.
- Number of layers: 12.
- Attention heads: 12.

To adjust the class to the specific usage, it was modified to classify two labels specifically for deepfake detection:

- Class 0: Original (Real) videos.
- Class 1: Manipulated (Fake) videos.

C. *Data Preprocessing*

To maintain consistency across datasets and prepare them for input into the models, each video undergoes a standardized data preparation process.

1) *Frame extraction and preprocessing*

To prepare the video data for the Vision Transformer model, each video underwent a frame extraction process. A total of 15 frames were uniformly extracted from each video to serve as a representative sample of its content. This process resulted in a total dataset of 30,000 image frames (2000 videos×15 frames/video), which were then converted to RGB format.

Every frame taken from the videos is first handled using the ViT feature extractor. This method regularizes pixel values and image size. Maintaining homogeneity throughout the input data helps guarantee that the data is compatible with the model and hence improves the efficiency of the training process.

The function of feature extraction turns the obtained frames into a tensor form fit for the model input. It also helps one to change image proportions to satisfy certain needs. The function uses a technique of normalizing to guarantee consistency.

2) *Dataset creation*

After processing each frame from every video—both the original and the altered versions—they are combined with their appropriate class labels (0 for the original version and 1 for the fake version). The last dataset produced by this procedure is then used in the stages of training and evaluation.

D. *Dataset Division*

The dataset of 30,000 frames was first partitioned into a training set (90%, or 27,000 frames) and a final hold-out test set (10%, or 3000 frames). The test set was kept separate and was only used for the final performance evaluation after the model was trained.

To ensure a robust and reliable evaluation of the model, a 5-fold cross-validation strategy was employed on the 27,000-frame training set. For each fold, the data was split into a training portion (4/5, approx. 21,600 images) and a validation portion (1/5, approx. 5400 images). This process was repeated five times, with each fold serving as the validation set once. This method provides a more stable estimate of the model's performance and minimizes the risk of selection bias from a single train-validation split.

E. *Model Refinement (Fine-Tuning)*

The pre-trained google/vit-base-patch16-224 model was fine-tuned using the following hyperparameters for each fold of the cross-validation process:

- Epochs per Fold: 3.
- Optimizer: AdamW.
- Learning Rate: $2e^{-5}$.
- Batch Size: 32.
- Weight Decay: 0.01.
- Hardware: NVIDIA A100 GPU.

F. *Evaluation Metrics*

The metrics under consideration are measured using the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) parameters, abbreviated hereafter. The following formula is employed to calculate the required parameters:

1) Accuracy

Measures the overall correct predictions across the test set, providing a general indication of model performance.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

2) Precision

The proportion of true positive predictions (the ability of the model not to mislabel the original video as fake).

$$Prec = \frac{TP}{TP + FP} \quad (2)$$

3) Recall

Proportion of actual positive cases correctly identified (ability of the model to find all false videos).

$$Rec = \frac{TP}{TP + FN} \quad (3)$$

4) F1-Score

The harmonic mean of Precision and Recall, provides a measure of the balance between the two.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

5) Confusion matrix

Used to evaluate the performance by displaying the number of true positive, true negative, false positive, and false negative predictions. It helps to assess the model's accuracy and identify areas for improvement.

IV. Result and Discussion

In the following section, the results of the refinement and evaluation of the Vision Transformer model are presented. The analysis is divided into two main stages: firstly, the evaluation of the model stability through a 5-fold cross-validation process, and secondly, the final performance evaluation on a hold-out test set to measure the generalisation ability of the model on data that has not been seen before.

A. Cross-Validation Performance

To ensure that the performance of the model does not depend on the random distribution of data, we applied a 5-fold cross-validation strategy to the training dataset. The performance metrics (Accuracy, Precision, Recall, and F1-Score) were calculated for each fold.

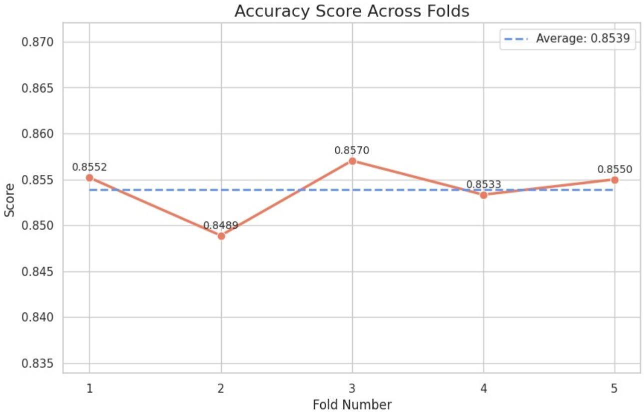

Fig. 5. Accuracy score across folds.

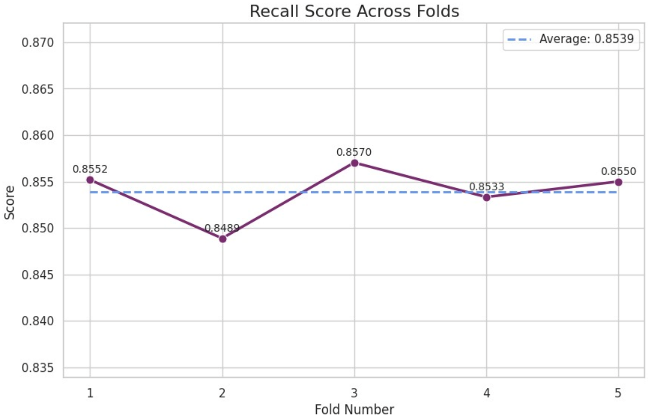

Fig. 6. Recall score across folds.

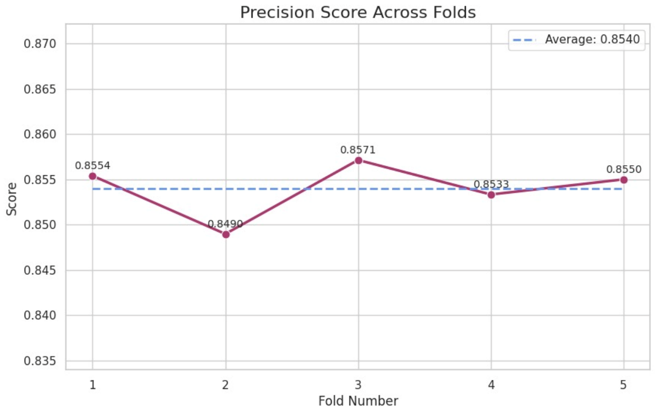

Fig. 7. Precision score across folds.

Fig. 5 shows the accuracy for each fold during the 5-fold validation process. The dashed blue line shows the average performance of the five folds used which is 0.8539. The use of 5-fold cross validation shows that the proposed model is robust, this is shown by the consistent accuracy results for each iteration.

The recall metric (also known as sensitivity or True Positive Rate) is crucial in deepfake detection tasks. A high recall indicates that the model is able to minimise the number of False Negatives. The average recall value during the 5-fold run was 0.8539. This proves that the model has a consistent ability to select deepfake content, as illustrated in Fig. 6.

In addition to having adequate sensitivity (high recall), the model also has a high level of confidence (high precision). The stable average precision score of 0.8540 (Fig. 7) shows that the model is reliable and does not make false accusations frequently.

The consistently high precision, precision and recall, through multiple iterations, indicate that the model effectively learns and performs well in distinguishing between the original and manipulated samples. Small improvements indicate that the model continues to learn without significant overfitting.

B. *Evaluation Metrics*

To assess the performance of the fine-tuned version of the Google's ViT-Base-Patch16-224 model for detecting deepfake, standard classification evaluation metrics were employed, including precision, recall, and F1-Score. The results are summarized in the classification report below:

TABLE I. CLASSIFICATION REPORT

Index	Precision	Recall	F1-Score	Support
0	0.8415	0.8713	0.8562	1500
1	0.8666	0.8360	0.8510	1500
Accuracy		0.8536		
Macro Average	0.8541	0.8536	0.8536	3000
Weighted Average	0.8541	0.8536	0.8536	3000

After the cross-validation process, the model was evaluated one last time using a hold-out test set consisting of 3000 images. The results of this test can be seen in Table I. The classification report indicates that the model achieved an overall accuracy of 0.8537. Most significant is the balanced performance between the two classes. The "Real" class (0) achieved a recall of 0.871 and F1-Score of 0.856, while the "Fake" class (1) achieved a recall of 0.836 and F1-Score of 0.851. These results prove that with a balanced dataset, the ViT model is able to learn distinguishing features for both classes effectively.

C. *Confusion Matrix*

The confusion matrix evaluation, visualised in Fig. 8, further highlights the performance of the model.

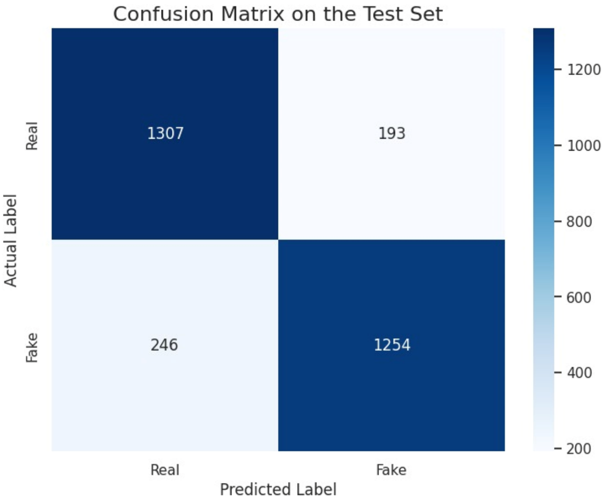

Fig. 8. Confusion matrix visualization.

The number can be interpreted as below:

- True Negatives (1307): The model correctly identified 1,307 real videos.
- True Positives (1254): The model correctly identified 1,254 fake videos.
- False Negatives (193): The model incorrectly identified 193 fake videos as real videos.
- False Positives (246): The model incorrectly identified 246 real videos as fake videos.

The relatively balanced number of False Positives and False Negatives indicates no significant bias towards either class, confirming the findings of the classification report. In the data considered as False Positive, it was found that most of them had poor visual quality such as high video compression or lack of lighting (examples are shown in Fig. 9). Furthermore, some False Negatives are high-quality deepfake videos where the manipulation is very subtle as in the face (an example is shown in Fig. 10).

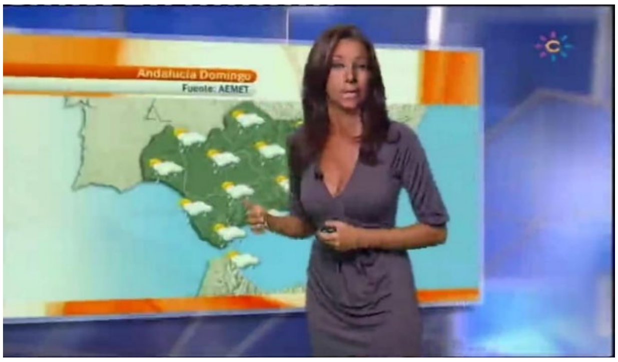

Fig. 9. Example of data that is considered false positive.

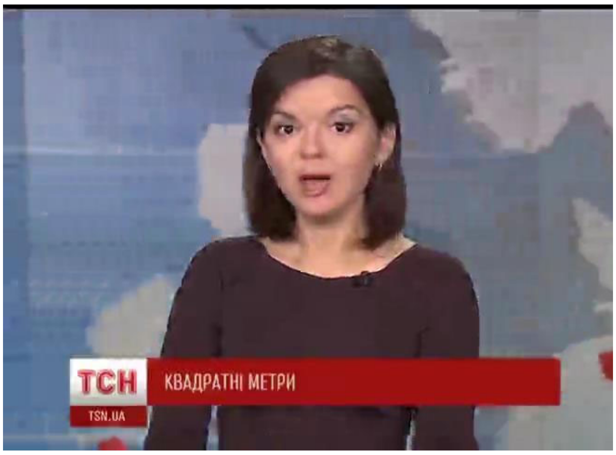

Fig. 10. Example of data that is considered false negative.

This experiment successfully demonstrates that the Vision Transformer model fine-tuned on a balanced dataset is capable of being a reliable deepfake detection tool. The accuracy of 85.37% that was rigorously validated through cross-validation is a robust and reliable result.

To give context to these results, we refer to other studies that utilise CNN architectures. For example, Ashok and Joy [21] reported that XceptionNet was able to record 95% accuracy on uncompressed and high-quality videos, and more than 80% accuracy on highly compressed videos. It should be noted that our results cannot be directly compared with theirs due to fundamental differences in evaluation methodology (our 5-fold cross-validation vs. their single train-test split). However, the stable performance we demonstrate under this more rigorous evaluation framework underlines the potential of ViT as a robust alternative for deepfake detection.

V. Conclusion

Deepfakes can have a significant negative impact if not handled properly. Many methods and algorithms can be used to deal with the problem. In this study, by applying a more robust methodology-using a balanced FaceForensics++ dataset (1000 real videos and 1000 fake videos) and a 5-fold cross-validation strategy-we managed to overcome the data bias problem that often plagues similar studies. Based on the experimental results, the ViT model delivers consistent performance. The cross-validation process demonstrates this with an average accuracy of 85.39%.

Despite the promising results, we recognise some limitations. Our current approach is still based on per-frame analysis and does not utilise the temporal information contained in the video. This study effectively employed a balanced subset to guarantee impartial training; however, a crucial direction for future research is to assess the model's performance on the complete, imbalanced FaceForensics++ dataset. To see how strong and useful the model is in real-world situations where data is often skewed, you would need to try out and compare different ways of managing data, such as SMOTE or resampling techniques. Finally, the evaluation is based on only one dataset and has not been validated against more advanced and recent manipulation methods.

Ethical Statement

In the preparation of this manuscript, the authors utilized a generative artificial intelligence tool (Quillbot) as a language editing assistant. The role of the AI was strictly limited to improving grammar, refining sentence structure, and enhancing the overall clarity and academic tone of the text.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Orvis Lutanora Siagian: Writing—Original draft preparation, Data Curation, Conceptualization, Methodology, Software, Formal analysis. Reinhard Ebenhaizer: Writing—Original draft preparation, Data Curation, Conceptualization, Methodology. Pandu Wicaksono: Methodology, Supervision, Project administration, Funding acquisition, Review & Editing. Zahra Nabila Izdihar: Methodology, Supervision, Funding acquisition, Review. All authors had approved the final version.

References

[1] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention," *IEEE Access*, vol. 12, pp. 8932–8939, 2024. https://api.semanticscholar.org/CorpusID:266969401

[2] Y. J. Heo, W. H. Yeo, and B. G. Kim, "DeepFake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527, 2022. https://api.semanticscholar.org/CorpusID:251116857

[3] S. Alanazi and S. Asif, "Exploring deepfake technology: Creation, consequences and countermeasures," *Human-Intelligent Systems Integration*, vol. 6, no. 1, pp. 49–60, 2024. https://api.semanticscholar.org/CorpusID:272732210

[4] B. U. Mahmud and A. A. Sharmin, "Deep insights of deepfake technology: A review," arXiv Print, arXiv:2105.00192, 2021. doi.org/10.48550/arXiv.2105.00192

[5] M. Pawelec, "Decent deepfakes? Professional deepfake developers' ethical considerations and their governance potential," *AI and Ethics*, vol. 5, no. 3, pp. 2641–2666, 2024. https://api.semanticscholar.org/CorpusID:272888269

[6] S. A. Hussien and S. N. Mohamed, "DeepFake video detection using vision transformer," *International Journal of Intelligent Computing and Information Sciences*, vol. 24, no. 1, pp. 55–68, 2024. https://api.semanticscholar.org/CorpusID:268701920

[7] B. K. Ruan, H. H. Shuai, and W. H. Cheng, "Vision transformers: State of the art and research challenges," arXiv Print, arXiv:2207.03041, 2022. doi.org/10.48550/arXiv.2207.03041

[8] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake image detection using vision transformer models," in *Proc. 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, 2024, pp. 332–335. https://api.semanticscholar.org/CorpusID:272355208

[9] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022. https://api.semanticscholar.org/CorpusID:247116742

[10] L. Payne. (2023). Deepfake history & facts britannica. [Online]. Available: https://www.britannica.com/technology/deepfake

[11] F. Siddiqui, J. Yang, S. Xiao, and M. Fahad, "Enhanced deepfake detection with DenseNet and Cross-ViT," *Expert Systems with Applications*, vol. 267, 126150, 2025. https://doi.org/10.1016/j.eswa.2024.126150

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. J'egou, "Training data-efficient image transformers & distillation through attention," in *Proc. International Conference on Machine Learning*, 2020, pp. 10347–10357. https://api.semanticscholar.org/CorpusID:229363322

[13] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2022, pp. 2286–2296, 2021. https://api.semanticscholar.org/CorpusID:232290742

[14] D. Wodajo, P. Lambert, G. V. Wallendael, S. Atnafu, and H. Mareen, "Improved deepfake video detection using convolutional vision transformer," in *Proc. 2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024, pp. 1–6. https://api.semanticscholar.org/CorpusID:271115309

[15] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Exploring self-supervised vision transformers for deepfake detection: A comparative analysis," in *Proc. 2024 IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–10. https:// api.semanticscholar.org/CorpusID:269484446

[16] C. Kumar, "Deepfake detection using parallel vision transformers," *Computer Science,* 2024. https://api.semanticscholar.org/CorpusID:277410873

[17] M. A. Arshed, A. S. Alwadain, R. F. Ali, S. Mumtaz, M. Ibrahim, and A. Muneer, "Unmasking deception: Empowering deepfake detection with vision transformer network," *Mathematics*, vol. 11, no. 17, p. 3710, 2023. https://api.semanticscholar.org/CorpusID:261391802

[18] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1831–1839. https://api.semanticscholar.org/CorpusID:4533859

[19] Y. J. Heo, Y. J. Choi, Y. W. Lee, and B. G. Kim, "Deepfake detection scheme based on vision transformer and distillation," arXiv Print, arXiv:2104.01353, 2021. doi.org/10.48550/arXiv.2104.01353

[20] M. Sharafudeen, and V. Chandra SS, "Leveraging vision attention transformers for detection of artificially synthesized dermoscopic lesion deepfakes using Derm-CGAN," *Diagnostics*, vol. 13, no. 5, p. 825, 2023. https://api.semanticscholar.org/CorpusID:257132942

[21] A. Ashok and P. T. Joy, "Deepfake detection using XceptionNet," in *Proc. 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 2023, pp. 1–5. https://api.semanticscholar.org/CorpusID:266563493

[22] Y. M. B. G. Cunha, B. R. Gomes, J. M. C. Boaro *et al*., "Learning self-distilled features for facial deepfake detection using visual foundation models: General results and demographic analysis," *J. Interact. Syst*., vol. 15, no. 1, pp. 682–694, 2024. https://api.semanticscholar.org/CorpusID:271284814

[23] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6. https://api.semanticscholar.org/CorpusID:61808533

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11. https://api.semanticscholar.org/CorpusID:59292011

[25] Google. (2020). *Google/vit-base-patch16-224 Hugging Face*. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224

[26] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," arXiv Print, arXiv:2104.10972, 2021. doi.org/10.48550/arXiv.2104.10972

[27] O. Russakovsky, J. Deng, H. Su *et al*., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2014. https://api.semanticscholar.org/CorpusID:2930547

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC-BY-4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.