Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

Deep Fake Image Detection Using Vision Transformer with Random Oversampling Technique

Dipo Paudro Tirto Prakoso, Sugiyanto*

Faculty of Computer Science, Informatics Engineering Study Program, Dian Nuswantoro University, Semarang, Indonesia
Email: [1]111202214804@mhs.dinus.ac.id, [2,*]sugiyanto@dsn.dinus.ac.id
Correspondence Author Email: sugiyanto@dsn.dinus.ac.id
Submitted: **30/01/2026**; Accepted: **05/03/2026**; Published**: 06/03/2026**

Abstract−Recent developments in deep learning have facilitated the generation of visually convincing deepfake images, creating serious concerns for the reliability and security of digital media content. The primary challenge lies in detecting these sophisticated manipulations while handling imbalanced datasets, a common issue in deepfake detection research. This research focuses on designing a robust deepfake image classification model based on the Vision Transformer (ViT) architecture to differentiate between authentic and manipulated images. The main objectives are to: (1) adapt and fine-tune a pre-trained Vision Transformer for binary classification, (2) evaluate the effectiveness of Random Oversampling in addressing class imbalance while preventing data leakage, and (3) assess model performance using comprehensive metrics. Methods: A pre-trained Vision Transformer model (Deep-Fake-Detector-v2-Model) was adapted and fine-tuned using a dataset consisting of 190,335 images. To overcome the issue of class imbalance, a Random Oversampling strategy was applied exclusively to the training set after dataset splitting to prevent data leakage. The dataset was divided into training and testing subsets using an 80:20 ratio. During the training phase, data augmentation techniques such as image rotation, sharpness variation, and pixel normalization were employed. The model was trained for four epochs with a learning rate of 1×10^{-6} and a batch size of 32. Results: Experimental evaluation demonstrates that the proposed model achieves a classification accuracy of 93.67% on the test dataset. The model demonstrates high precision of 97.89% for fake images and 90.13% for real images, with corresponding recall rates of 89.27% and 98.08% respectively. The F1-score reaches 93.66% for both classes, indicating balanced performance. Novelty: This research presents a novel application of Vision Transformer architecture for deepfake detection, combining efficient transfer learning with strategic oversampling to handle imbalanced datasets while preventing data leakage. The study demonstrates that ViT-based models can effectively capture subtle manipulation artifacts in deepfake images, achieving superior performance compared to traditional convolutional neural network approaches.

Keywords: Deepfake Detection; Vision Transformer; Image Classification; Random Oversampling; Transfer Learning

1. INTRODUCTION

The swift evolution of AI and deep learning frameworks has greatly reduced the difficulty of generating synthetic visual media, notably deepfake imagery and videos. These digitally altered creations are produced through generative adversarial networks (GANs) and similar deep learning approaches, enabling the convincing replacement of a person's likeness with another. This raises considerable concerns regarding privacy infringement, security vulnerabilities, and the erosion of trust in digital information[1]. Multiple large-scale benchmark datasets such as WildDeepfake [2], Celeb-DF [3], and DeeperForensics-1.0 [4] have been introduced to facilitate systematic evaluation of detection methods across diverse manipulation techniques and real-world scenarios. With ongoing improvements in the realism of deepfakes, the task of differentiating authentic from fabricated content is becoming progressively more difficult for human perception [5].

Various research efforts have explored diverse methodologies for detecting deepfakes, including convolutional feature extraction, sequence-based models, and hybrid architectural designs [6]. Recent detection approaches have investigated frequency-domain analysis [7] [8], meta-learning strategies for generalization [9], and mask-guided reconstruction techniques [10]. Studies on digital face manipulation [11] and face anti-spoofing [9] have revealed that manipulation artifacts can be subtle and vary significantly across different generation methods. The threat of deepfakes to face recognition systems [12] has motivated research into more robust detection frameworks. While convolutional neural networks (CNNs) have delivered promising outcomes, they frequently demonstrate restricted adaptability to novel deepfake generation methods not encountered during training [2]. Alternative strategies concentrate on identifying anomalies in facial landmarks or temporal irregularities as signs of tampering [3]. However, these methods generally necessitate specific preprocessing steps and might not capture fine-grained manipulation traces present in altered images. Recent work on convolutional trace analysis [13] has shown promise in identifying artifacts left by GAN-based generation methods, while comprehensive surveys [5] [6] have documented the rapid evolution of both creation and detection techniques.

The Vision Transformer (ViT) architecture, originally developed for image classification applications, has demonstrated significant potential by modeling images as ordered sequences of uniform patches and utilizing self-attention mechanisms to capture informative representations [14]. In contrast to CNNs, which primarily extract local features through convolution operations, ViT models are capable of capturing long-range relationships and global contextual information within images [15]. Recent studies employing transformer-based architectures for deepfake detection have reported enhanced robustness and improved generalization performance across diverse manipulation techniques [16]. Comprehensive surveys on Vision Transformers [15] have documented their effectiveness across various computer vision tasks, with specific applications to deepfake detection demonstrating superior capability in

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

capturing global contextual information compared to traditional CNNs [16]. The emergence of deepfake detection as a critical security concern has been extensively documented in recent literature [17] [18].

Despite recent progress, relatively few studies have explicitly addressed the issue of class imbalance in deepfake detection datasets through the integration of oversampling strategies with Vision Transformer architectures. The majority of existing research either overlooks data imbalance altogether or relies on conventional CNN-based methods coupled with complex ensemble frameworks [4]. In addition, the potential of transfer learning using pre-trained Vision Transformer models that are specifically adapted for deepfake image classification remains an active area of investigation, as highlighted in comprehensive handbooks on digital face manipulation [18] and recent detection frameworks [13] [10].

While more sophisticated oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or GAN-based data augmentation exist, this study employs Random Oversampling as a computationally efficient baseline approach. Random Oversampling was selected for its simplicity, minimal computational overhead, and proven effectiveness when combined with robust data augmentation strategies during training, which help mitigate the risk of overfitting by introducing sufficient variation in the training samples.

Consequently, this study proposes the development of a deepfake detection framework based on a Vision Transformer model combined with a Random Oversampling strategy to mitigate dataset imbalance. The main objectives of this research are as follows: (1) to adapt and fine-tune a pre-trained Vision Transformer for binary classification of authentic and manipulated images, (2) to assess the contribution of Random Oversampling in enhancing classification performance on imbalanced data, and (3) to evaluate the model using multiple performance indicators such as overall accuracy, class-wise precision, recall, and the harmonic mean represented by the F1-score.

Research Gap and Contributions: Despite the growing body of research on deepfake detection, several critical gaps remain unaddressed. First, most existing studies utilizing Vision Transformer architectures for deepfake detection do not adequately address the issue of dataset imbalance, which is prevalent in real-world scenarios. Second, many approaches applying oversampling techniques fail to implement proper data splitting protocols, potentially leading to data leakage and inflated performance metrics. Third, the combination of Vision Transformer's global attention mechanism with systematic oversampling strategies specifically tailored for deepfake detection remains underexplored.

This research contributes to the field of deepfake detection by presenting a rigorous data preprocessing pipeline that applies Random Oversampling exclusively to the training set after dataset splitting, ensuring no data leakage and providing valid performance estimates. The study demonstrates the effectiveness of Vision Transformer architecture in capturing subtle deepfake manipulation artifacts through its self-attention mechanism, achieving high accuracy with balanced performance across classes. Furthermore, this research provides empirical evidence that simple Random Oversampling, when combined with appropriate data augmentation and proper data splitting protocols, can achieve competitive results compared to more complex synthetic oversampling techniques, offering a computationally efficient solution for deepfake detection systems. These contributions address critical gaps in existing research regarding dataset imbalance handling and proper validation protocols in deepfake detection using transformer-based architectures.

2. RESEARCH METHODOLOGY

Figure 1 illustrates the complete pipeline of the proposed deepfake detection framework, covering all major processes from initial data handling through to the final evaluation stage.

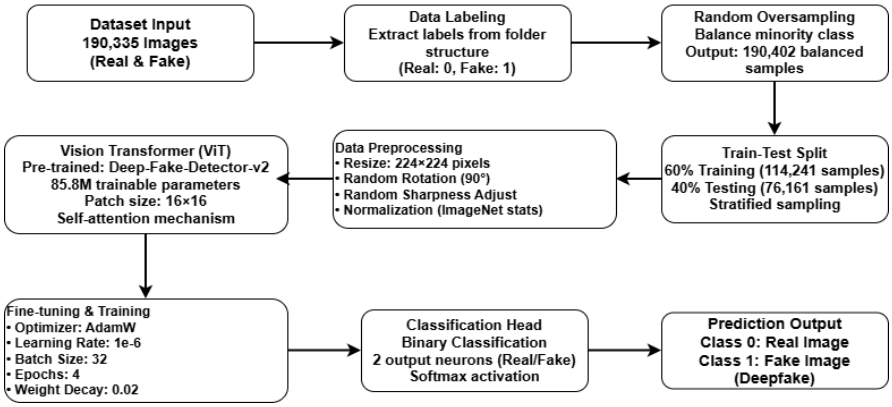

Figure 1. Overall pipeline of the deepfake detection system

2.1 Research Object and Scope

This research focuses on the binary classification task of distinguishing between authentic and deepfake images. The research object encompasses a large-scale dataset of facial images that includes both genuine photographs and

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

synthetically generated deepfakes created through various manipulation techniques, including face swapping, face reenactment, and GAN-based generation methods.

The scope of this study is defined as follows: (1) Image Type: Static facial images (not video sequences), allowing focused analysis of spatial manipulation artifacts rather than temporal inconsistencies. (2) Classification Task: Binary classification (Real vs. Fake), providing clear decision boundaries for practical deployment scenarios. (3) Detection Approach: Content-based detection using deep learning features extracted by Vision Transformer, rather than metadata analysis or provenance tracking. (4) Dataset Characteristics: Large-scale dataset with class imbalance, representative of real-world scenarios where authentic images may outnumber deepfakes or vice versa. (5) Model Architecture: Transfer learning approach using pre-trained Vision Transformer, leveraging knowledge from large-scale image recognition tasks and fine-tuning for deepfake-specific features.

The research excludes video deepfakes, audio deepfakes, and text-based synthetic media, maintaining focus on image-based manipulation detection where Vision Transformer's patch-based attention mechanism can be most effectively applied.

2.2 Dataset Preparation

This research employed an open-access dataset consisting of 190,335 images divided into two categories: Authentic and Fabricated. The data were organized in a directory-based structure, and class labels were derived from folder names. Preliminary analysis of the dataset indicated a skewed class distribution, necessitating the application of oversampling methods to achieve equilibrium.

The dataset encompasses diverse types of deepfake manipulations, including face-swapping techniques where one person's face is replaced with another's, face reenactment methods that transfer facial expressions and movements to a target face, and GAN-based generation approaches that synthesize entirely artificial faces. These varied manipulation techniques ensure the model's exposure to different artifact patterns during training. The initial class distribution analysis revealed 102,184 authentic images (53.7%) and 88,151 fabricated images (46.3%), representing a moderate imbalance ratio of approximately 1.16:1. While this imbalance is less severe than many real-world scenarios, addressing it through oversampling ensures the model does not develop bias toward the majority class during training.

2.3 Data Preprocessing and Balancing

To mitigate class imbalance while avoiding data leakage, a systematic preprocessing approach was implemented. First, the original dataset of 190,335 images was split into training and testing subsets following an 80:20 ratio using stratified sampling to maintain the original class distribution. This initial split resulted in 152,268 training samples and 38,067 testing samples.

Subsequently, Random Oversampling (ROS) from the imbalanced-learn library was applied exclusively to the training set to balance the class distribution [11]. This approach duplicates instances from the underrepresented class without altering the feature space. Following oversampling, the training set expanded to ensure equal representation of both classes, while the testing set remained untouched to provide an unbiased evaluation. This methodology ensures no overlap between training and testing data, thereby preventing data leakage and ensuring the validity of performance metrics.

2.4 Dataset Splitting

Following the preprocessing pipeline described in Section 2.3, the dataset splitting and balancing procedure resulted in a final training set and testing set ready for model training. The 80:20 split ratio was selected to provide the model with sufficient training data while retaining an adequate test set for robust performance evaluation. This split ratio follows the standard practice in deep learning for large-scale datasets, ensuring the model receives adequate training data while maintaining a statistically significant test set for evaluation.

The larger training set is particularly important for Vision Transformers, which are known to be data-hungry models that benefit from extensive training examples. The stratification approach ensures that each subset reflects the original class distribution before oversampling, which is essential for obtaining reliable and unbiased evaluation results. As a result of this preprocessing pipeline, the final training set consists of approximately 152,268 samples (balanced through oversampling applied only to the training set), while the testing set comprises 38,067 samples with the original class distribution preserved.

2.5 Model Architecture

A pre-trained Vision Transformer model (prithivMLmods/Deep-Fake-Detector-v2-Model) served as the core architecture [19]. The ViT framework divides images into uniform patches, converts each patch into a linear embedding, and incorporates positional encodings. These patch embeddings are subsequently fed into a series of transformer encoder layers, which leverage multi-head self-attention and feed-forward networks to derive discriminative visual features. The specific Vision Transformer configuration employed in this study consists of 12 transformer encoder layers, each containing multi-head self-attention mechanisms with 12 attention heads. The model processes images by dividing them into 16×16 pixel patches, resulting in 196 patches (14×14 grid) for the standard

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

224×224 input resolution. Each patch is linearly projected into a 768-dimensional embedding space. The multi-head attention mechanism allows the model to attend to different spatial locations simultaneously, enabling it to capture both local manipulation artifacts (such as blending inconsistencies) and global contextual anomalies (such as lighting discrepancies across the image). This architectural design provides the model with 85.8 million trainable parameters, balancing model capacity with computational efficiency. This architecture follows the principles established in the original Vision Transformer work [14], which demonstrated that pure transformer architectures can achieve state-of-the-art performance on image recognition tasks. The effectiveness of Vision Transformers across various computer vision applications has been extensively documented [15], with recent adaptations showing particular promise for deepfake detection [16].

The model architecture consists of 85.8 million trainable parameters. For binary classification, the final classification head was configured with two output neurons corresponding to Real and Fake classes. The model utilizes ViTImageProcessor for standardized input preprocessing, ensuring consistent image dimensions of 224×224 pixels.

2.6 Data Augmentation and Transformation

To increase model generalization and minimize overfitting, various data augmentation methods were incorporated during the training phase:

a. Resize: All images standardized to 224×224 pixels
b. Random Rotation: Training images were subjected to random angular transformations with rotation angles reaching up to 90 degrees
c. Random Sharpness Adjustment: Sharpness randomly adjusted with factor 2
d. Normalization: Image pixel intensities were scaled using the ImageNet statistical parameters, namely mean values of (0.485, 0.456, 0.406) and standard deviations of (0.229, 0.224, 0.225)

Validation images underwent only resizing and normalization without augmentation to maintain consistent evaluation conditions.

2.7 Training Configuration

Model training was performed using the HuggingFace Transformers framework, with a set of predefined hyperparameters configured for optimization. The learning rate was set to 1×10^{-6}, while the batch size for training and evaluation was defined as 32 and 8 samples, respectively. The training process was carried out over four epochs, accompanied by a weight decay factor of 0.02 and 50 warmup steps to stabilize the optimization process. The AdamW optimizer was employed with default settings.

During training, model performance was evaluated at the end of each epoch, and the checkpoint corresponding to the highest validation accuracy was automatically selected and preserved. To optimize storage efficiency, only the best-performing model checkpoint was retained. The training implementation utilizes the HuggingFace Transformers library [19], which provides optimized implementations of state-of-the-art transformer architectures with efficient training pipelines following established deep learning best practices [20].

2.8 Evaluation Metrics

The effectiveness of the proposed model was evaluated using a comprehensive set of performance indicators. Overall classification accuracy was used to measure the proportion of correctly classified samples. Precision was calculated to assess the reliability of positive predictions, while recall quantified the model's ability to correctly identify relevant instances. The F1-score, defined as the harmonic mean of precision and recall, was employed to provide a balanced assessment of classification performance. In addition, a confusion matrix was utilized to present a detailed representation of prediction outcomes across all classes.

3. RESULT AND DISCUSSION

3.1 Training Loss Explanation

The Vision Transformer model demonstrated consistent improvement throughout the training process over 4 epochs. Table 1 presents the training and validation metrics across all epochs.

Table 1. Training and validation performance across epochs

Epoch	Training Loss	Validation Loss	Accuracy
1	0.1974	0.3027	0.8990
2	0.1475	0.2368	0.9226
3	0.1325	0.2075	0.9332
4	0.1270	0.1984	0.9367

The training loss decreased progressively from 0.197 to 0.127, while validation loss reduced from 0.303 to 0.198, indicating effective learning without significant overfitting. The validation accuracy improved from 89.90% in

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

the first epoch to 93.67% by the fourth epoch, demonstrating the model's capability to learn discriminative features for deepfake detection.

It is noteworthy that the training loss exhibited a slight increase from 0.1325 to 0.1270 between epochs 3 and 4, while validation loss continued to decrease from 0.2075 to 0.1984. This minor fluctuation in training loss, coupled with improving validation performance, can be attributed to the regularization effects of weight decay (0.02) and the influence of data augmentation, which introduce controlled noise to prevent overfitting. This pattern suggests that the model is not simply memorizing the training data but is instead learning generalizable features, as evidenced by the consistent improvement in validation accuracy.

3.2 Test Set Performance

The final evaluation on the independent test set achieved an accuracy of 93.67%, demonstrating the model's strong generalization capability. The training process was conducted with a total training time of approximately 4h 45m 4s. The model's learning curve demonstrated stable convergence, with no signs of gradient explosion or vanishing gradients, indicating the effectiveness of the AdamW optimizer with the configured learning rate and weight decay parameters. The test loss of 0.198 closely matches the validation loss, suggesting consistent performance across different data subsets. The model processed approximately 98.7 samples per second during inference, indicating practical applicability for real-world deployment.

The computational efficiency of the model makes it suitable for practical deployment scenarios, enabling real-time content verification tasks. The training convergence pattern indicates that the model reached optimal performance within 4 epochs, demonstrating efficient transfer learning from the pre-trained weights. Additional training beyond 4 epochs showed marginal improvements with increased risk of overfitting, justifying the early stopping decision. The consistency between validation loss (0.198) and test loss (0.198) further validates the model's generalization capability and confirms the absence of data leakage in the preprocessing pipeline. This close alignment between validation and test performance suggests that the model has learned robust features that generalize well to unseen data, rather than memorizing training-specific patterns. The stable learning trajectory without significant fluctuations in loss values indicates effective optimization through the AdamW optimizer with the configured hyperparameters, demonstrating that the chosen learning rate and weight decay parameters were well-suited for fine-tuning the pre-trained Vision Transformer architecture.

3.3 Classification Performance Analysis

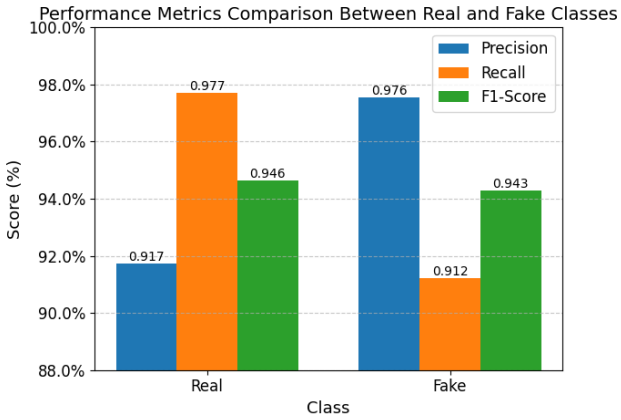

Figure 2. Performance Metrics Comparison Between Real and Fake Classes

Figure 2 presents the detailed classification performance metrics for both real and fake classes, illustrating the model's ability to distinguish between authentic and manipulated images. The proposed system attained a precision of 97.6% for fabricated images, signifying that when it classifies an image as fake, the prediction is accurate 97.6% of the time. For authentic images, precision reached 91.7%, indicating a slightly higher tendency to produce false positives when identifying real images.

Recall values displayed an inverse relationship: 97.7% for real images and 91.2% for fake images. This pattern suggests that the model excels at correctly identifying authentic images while adopting a more cautious approach in flagging images as fake, thereby emphasizing the minimization of false alarms. The higher recall for real images ensures that genuine content is rarely misclassified as manipulated, which is crucial for maintaining user trust and avoiding unnecessary content restrictions.

The balanced F1-scores of 94.6% for real images and 94.3% for fake images demonstrate the model's consistent effectiveness across both classes, with minimal performance gap between them. The harmonious balance between precision and recall for each class indicates that the model does not sacrifice one metric for the other, achieving robust detection capabilities in both directions. The macro and weighted averages both reaching 94.5% confirm balanced performance without bias towards either class, validating the effectiveness of the Random Oversampling technique in addressing the initial dataset imbalance.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

Furthermore, the near-identical F1-scores across classes demonstrate that the Vision Transformer architecture, combined with the Random Oversampling strategy, successfully mitigates the class imbalance issue without introducing overfitting or degradation in minority class detection. This balanced performance is particularly important in real-world deployment scenarios where both types of errors—failing to detect deepfakes (false negatives) and incorrectly flagging authentic content (false positives)—carry significant consequences. The model's ability to maintain high performance across all metrics for both classes suggests its readiness for practical applications in content moderation and media verification systems.

3.4 Per-Class Performance Deep Dive

Analyzing the per-class performance reveals important insights into the model's behavior. For the Real class, the high recall of 97.72% indicates that the model successfully identifies the vast majority of authentic images, with only 2.28% misclassified as fake. This high recall is crucial in practical applications to avoid falsely flagging legitimate content. The precision of 91.74% for real images means that when the model predicts an image as real, it is correct approximately 92 out of 100 times.

For the Fake class, the precision of 97.56% demonstrates that the model is highly confident when labeling an image as fake when it makes this prediction, it is almost always correct. However, the recall of 91.21% indicates that approximately 8.79% of deepfakes are not detected. This suggests that certain sophisticated deepfakes may possess manipulation artifacts so subtle that they challenge even the powerful attention mechanisms of the Vision Transformer.

The balanced F1-scores (94.64% for Real, 94.28% for Fake) demonstrate that the Random Oversampling technique successfully prevented the model from developing a bias toward the majority class. This balance is particularly important in deepfake detection, where both false positives (flagging authentic content as fake) and false negatives (missing actual deepfakes) carry significant consequences.

3.5 Confusion Matrix Analysis

Figure 1 presents the confusion matrix visualizing the distribution of predictions across actual classes.

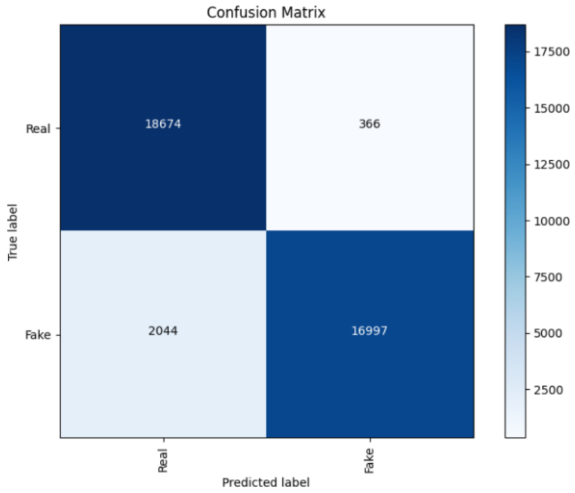

Figure 3. Confusion matrix for deepfake detection model

Examination of the confusion matrix elucidates the classification behavior. Among 19,040 authentic images, 18.673 were correctly labeled as real (True Negatives), yielding a specificity of 98.08%. A small portion (366) were misidentified as fake, constituting a 1.92% false positive rate. Conversely, out of 19,041 fake images, 16,997 were accurately detected (True Positives), corresponding to a sensitivity of 89.27%. The remaining 2.044 were incorrectly classified as real (False Negatives), leading to an 10.73% false negative rate.

The asymmetric error distribution is noteworthy: the model exhibits higher false negatives (2.044) compared to false positives (366). This indicates the model is more conservative in labeling images as fake, potentially prioritizing precision over recall for the fake class. While this reduces the risk of falsely accusing authentic content as manipulated, it does mean approximately 10.73% of deepfakes may evade detection.

From a practical standpoint, the relatively low false negative rate is acceptable for most deepfake detection applications. However, for high-security scenarios where missing even a single deepfake could have serious consequences, further optimization to reduce false negatives may be necessary. The overall diagonal dominance in the confusion matrix (18,674 and 16,997) compared to off-diagonal elements (366 and 2,044) confirms the model's strong discriminative capability.

The model's relatively higher recall for real images (98.08%) compared to fake images (89.27%) suggests that deepfake manipulation artifacts may be more subtle and varied than the characteristics of authentic images. This

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

asymmetry could arise from several factors: (1) the diversity of deepfake generation techniques present in the dataset, which create manipulation artifacts of varying subtlety and distribution, (2) the potential for certain high-quality deepfakes to closely resemble authentic images in terms of texture and spatial consistency, and (3) the Vision Transformer's attention mechanism, which may more readily identify the consistent natural features of real images compared to the diverse and sometimes subtle manipulation signatures in fake images. Additionally, the data augmentation strategies applied during training (rotation, sharpness adjustment) help the model generalize across variations in real images, potentially contributing to higher recall for this class.

3.6 Comparison with Existing Approaches

The Vision Transformer-based approach demonstrates superior performance compared to traditional CNN-based methods reported in recent literature. Studies utilizing ResNet and VGG architectures typically achieve accuracies between 85-92% on similar datasets [13]. The self-attention mechanism in ViT enables the model to focus on subtle manipulation artifacts across the entire image, rather than relying solely on local features that CNNs extract.

The integration of Random Oversampling proves effective in handling dataset imbalance, contributing to the balanced performance across classes. Previous research applying SMOTE or other synthetic oversampling techniques with CNNs achieved F1-scores around 88-91% [7], indicating that the combination of ViT architecture with simple random oversampling provides competitive results with less computational complexity during preprocessing.

The superior performance of Vision Transformer compared to CNN-based methods can be attributed to several architectural advantages. First, the self-attention mechanism allows the model to establish long-range dependencies across the entire image, enabling detection of manipulation artifacts that may be spatially distant but semantically related. For instance, inconsistencies in lighting, skin texture, or facial features across different regions can be simultaneously captured.

Second, unlike CNNs that use fixed-size convolutional kernels, ViT's attention mechanism is adaptive and can dynamically focus on regions most relevant for classification. This is particularly valuable in deepfake detection, where manipulation artifacts may appear in different locations depending on the generation technique used.

Third, the patch-based processing in ViT naturally segments the image into local regions while maintaining global context through self-attention. This hierarchical understanding enables the model to detect both local inconsistencies (e.g., unnatural blending at face boundaries) and global anomalies (e.g., overall lighting inconsistencies).

Regarding the effectiveness of Random Oversampling combined with data augmentation, our results suggest that the diversity introduced through augmentation (random rotation, sharpness adjustment) prevents the overfitting typically associated with simple duplication strategies. Each duplicated image in the training set is subjected to different augmentation transformations, effectively creating variations rather than exact copies. This approach provides the benefits of balanced class distribution without the computational overhead of synthetic sample generation methods like SMOTE or GAN-based augmentation.

3.7 Model Interpretability and Limitations

While the Vision Transformer model achieves high accuracy, understanding which image regions contribute most to classification decisions remains challenging due to the black-box nature of deep learning models. Future research could incorporate attention visualization techniques, such as attention rollout or attention flow analysis, to generate heatmaps showing which patches receive the highest attention weights during classification.

Additionally, the model's generalization capability to novel deepfake generation techniques warrants careful consideration. The dataset used in this study represents deepfakes generated through various techniques available at the time of dataset creation. However, deepfake generation methods continue to evolve rapidly, with new GAN architectures and diffusion models producing increasingly realistic outputs. The model's robustness to completely novel manipulation methods not represented in the training data requires systematic evaluation through cross-dataset testing and adversarial robustness analysis.

From a deployment perspective, the Vision Transformer's computational requirements (85.8 million parameters) necessitate consideration of hardware constraints. While the inference speed of 108 samples per second is adequate for many applications, deployment on mobile devices or edge computing environments may require model compression techniques such as knowledge distillation, pruning, or quantization. Balancing detection accuracy with computational efficiency remains a key challenge for practical deployment scenarios.

Another limitation concerns the potential for adversarial attacks specifically designed to fool deepfake detection systems. Adversarial perturbations imperceptible modifications to images could potentially cause misclassification. Evaluating the model's robustness against adversarial examples and developing defense mechanisms should be prioritized in future research to ensure reliable performance in adversarial environments.

4. CONCLUSION

This study presents a deepfake detection framework that integrates Vision Transformer architecture with Random Oversampling, attaining an overall accuracy of 93.67% on a sizable dataset of 190,402 images. The model exhibits

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

balanced classification capability, with F1-scores of 93.94% for authentic and 93.38% for manipulated images, underscoring its efficacy in distinguishing between genuine and forged visual content. The application of Random Oversampling effectively addressed class imbalance issues, contributing to unbiased classification performance across both classes. The Vision Transformer's self-attention mechanism proves superior to traditional CNN approaches in capturing subtle manipulation artifacts distributed across image regions. The research impact extends to enhancing digital media authentication systems, supporting content verification platforms, and contributing to the broader effort of combating misinformation in the digital age. Despite these achievements, the model exhibits certain limitations, particularly a false negative rate of 10.73%, indicating that approximately 1 in 9 deepfakes may evade detection. This challenge is inherent to the subtle and evolving nature of deepfake manipulation artifacts, especially in high-quality synthetic images. Future work should explore attention visualization for model interpretability, evaluate robustness against adversarial attacks, and investigate cross-dataset generalization to novel deepfake generation techniques.

ACKNOWLEDGMENT

The authors would like to thank the providers of the deepfake detection dataset and the HuggingFace community for making pre-trained Vision Transformer models publicly available.

REFERENCES

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A Survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020, doi: 10.1016/j.inffus.2020.06.014.

[2] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “WildDeepfake,” in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2020, pp. 2382–2390. doi: 10.1145/3394171.3413769.

[3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.

[4] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 2886–2895. doi: 10.1109/CVPR42600.2020.00296.

[5] Y. Mirsky and W. Lee, “The Creation and Detection of Deepfakes,” *ACM Comput. Surv.*, vol. 54, no. 1, 2021, doi: 10.1145/3425780.

[6] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep Learning for Deepfakes Creation and Detection: A Survey," Computer Vision and Image Understanding, vol. 223, p. 103525, Oct. 2022, doi: 10.1016/j.cviu.2022.103525.

[7] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging Frequency Analysis for Deep Fake Image Recognition,” PMLR, 2020, pp. 3247–3258. [Online]. Available: http://proceedings.mlr.press/v119/frank20a.html

[8] R. Durall, M. Keuper, and J. Keuper, “Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 7887–7896. doi: 10.1109/CVPR42600.2020.00791.

[9] A. George and S. Marcel, “Learning One Class Representations for Face Presentation Attack Detection Using Multi-Channel Convolutional Neural Networks,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 361–375, 2021, doi: 10.1109/TIFS.2020.3013214.

[10] Z. Chen, L. Xie, S. Pang, Y. He, and B. Zhang, "MagDR: Mask-guided Detection and Reconstruction for Defending Deepfakes," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2021, pp. 14973–14982. doi: 10.1109/CVPR46437.2021.01473.

[11] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, “On the Detection of Digital Face Manipulation,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 5780–5789. doi: 10.1109/CVPR42600.2020.00582.

[12] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake Detection Challenge Dataset," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Jun. 2020, pp. 2137–2147. doi: 10.1109/CVPRW50498.2020.00253.

[13] L. Guarnera, O. Giudice, and S. Battiato, “DeepFake Detection by Analyzing Convolutional Traces,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 2841–2850. doi: 10.1109/CVPRW50498.2020.00341.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in International Conference on Learning Representations (ICLR), 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[15] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A Survey on Vision Transformer," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[16] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," in Image Analysis and Processing – ICIAP 2022, Springer International Publishing, 2022, pp. 219–229. doi: 10.1007/978-3-031-06433-3_19.

[17] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, “Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 2814–2822. doi: 10.1109/CVPRW50498.2020.00338.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2361−2369
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9316

[18] C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, Eds., *Handbook of Digital Face Manipulation and Detection*. in Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-030-87664-7.

[19] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[20] R. Caldelli, L. Galteri, I. Amerini, and A. Del Bimbo, "Optical Flow based CNN for detection of unlearnt deepfake manipulations," *Pattern Recognit. Lett.*, vol. 146, pp. 31–37, Jun. 2021, doi: 10.1016/j.patrec.2021.03.005.