Enhancing Deepfake Detection Using Attention Mechanisms in Image Classification

Kavita Sultanpure*, Prajakta Kumbhare, Jay Kumavat, Nidhi Karkera, Mandar Mahapure, Ameya Pathak
Information Technology, Vishwakarma Institute of Technology, Pune, India

* Corresponding author
doi: https://doi.org/10.21467/proceedings.7.6.42

Abstract

Deepfake technology, where artificial intelligence is used in order to generate highly realistic manipulated media, still is a severe threat to digital media authenticity. Detection of most deepfakes is really one of the most prominent computer vision tasks that we have at present. In this paper, we propose a specific deepfake detector. It is an attention and convolutional neural network (CNN). We employ an attention layer in our network to enable better salient detail representation along with trying to distinguish between forged and original images. Metrics including accuracy, precision, recall, and F1-score are measures used to determine the performance of the model after its training on a dataset of original and forged images. We additionally show the performance of the model by greatly improving the predictive outputs through face landmark detection and an image preprocessing pipeline.

Keywords: Deepfake, CNN, Attention Mechanism

1 Introduction

The AI and ML sectors have been transformed and impacted almost all industries, including media production and computer vision. Of all the impacts of these technologies, one of the most concerning innovations is the creation of deepfakes—false media, mainly video and images. Deepfakes are created through advanced means like Generative Adversarial Networks (GANs) and can convincingly change voice and movement and thus spread disinformation, impersonation, and criminal activity. Detection of deepfakes is important among industries like journalism, law enforcement, cybersecurity, and social media. Since the emergence of deepfake technology, simple algorithms or human verification are no longer sufficient. Hence, deep learning algorithms were implemented that detect the artifacts and discrepancies of deepfake media. The most successful technique of the recent past was to use convolutional neural networks (CNNs) to learn the visual patterns of the image and video and differentiate between original and generated media. A current deepfake detection method employs attention mechanisms in CNN models. The attention mechanism enhances model performance by focusing on important features, making it highly effective in deep learning models. CNNs often struggle to capture subtle anomalies in deepfake images. However, attention layers improve detection accuracy by emphasizing facial features, lighting inconsistencies, and unnatural movements.

In this work, we train our deepfake detection model using a dataset containing both real and deepfake images. The attention mechanism refines feature extraction, helping the network classify images more accurately. Our model performs binary classification, distinguishing between real and fake images by analysing learned features. To rigorously assess the model's performance, we conduct independent validation using accuracy, precision, recall, and F1-score as evaluation metrics. By leveraging attention mechanisms, this research aims to enhance existing deepfake detection methods and contribute to a more robust solution against evolving deepfake threats. Furthermore, this paper explores real-world applications of deepfake detection and discusses how detection systems can adapt to emerging challenges posed by deepfake technology.

2 Literature Review

Deepfake technology, driven by advanced machine learning methods, has evolved much further than its level in recent years, to the extent that it became possible to produce hyper-realistic but completely synthetic

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

media. Methods such as Generative Adversarial Networks (GANs) made deepfakes possible to allow end-users to have access to content manipulation such as images, videos, and audio, becoming harder to forge from original content. Although deepfake technology, with its immense potential for creativity, especially for film and entertainment, created tremendous expectations, the same created unprecedented concern regarding using it to propagate falsehoods for nefarious purposes. The capability to generate convincing content ushered in the horrors of false information, identity, defamation, as well as use in manipulating politics by governments, just to mention a few, thus unleashing an ever-growing threat on society. For their mitigation, detection of deepfakes therefore emerged as an important, pressing field of study that sought to ensure digital content integrity. Deepfake detection involves inconsistency or artifact checks introduced by producing misleading content. In images, for example, detection is mostly focused on recognizing visual discontinuities such as abnormal illumination conditions, pixel feature abnormality, or face feature inconsistency. In deepfakes with audio, tests are more interested in identifying subtlety inconsistency in voice quality, tone, and intonation, exhibiting unnatural tampering. Deepfake videos are a very difficult case with spatial features (like lip-sync or deformations on face features) as well as time-related patterns (like unnatural movement or between frames mismatch) analysis required. As technology for making deepfakes will become increasingly a driving force for research, outdated manual checks combined with rule-based algorithms can no longer be enough. Powerful, autonomous strategies with machine and deep learning mechanisms involvement are the ones needed today in order to effectively counter today's booming faked content threats.

A collection of advanced approaches has been documented in the case of deepfake detection, covering diverse aspects of media. Lai et al. [1] explained the GM-DF (Generative Model for Deepfake Detection) framework as a technique that improves generalization ability by training deepfake detection models on diverse data sets. With this, it allows the detection model to become adaptable to more prevalent and adaptive features, improving generalization against unique types of deepfakes. Rossler et al. [2] came up with the FaceForensics++ dataset and paired detection model designed for the detection of manipulations in faces from video. Drawing motivation from large collections of heterogeneous training data, they greatly improve video-based facial feature examination accuracy as a primary subsystem of video deepfake detection. MesoNet et al. [3], however, proposes a shallow video deepfake detection approach through effective feature representation of deepfakes with light computational loads and is thus easy to use for real-time detection.

Durall et al. [4] ventured deeper into detection using frequency domain analysis, a paradigm that explores underground frequency-based artifacts underlying deepfake multimedia. Their algorithm works well with even with limited number of labeled samples, with high potential in the limited training data scenario. Coccomini et al. [5] proposed Vision Transformers with EfficientNet, synergistic combination of best practices of individual models to capture spatial-temporal dependencies better in video deepfakes with better detection accuracies in visually dynamic content. Heidari et al. [6] proposed extensive review of number of deepfake detection approaches which are categorized in terms of classification as image, video, audio, and multimodal approaches. Their work reveals the path of deepfake detection approaches and identifies the need to come up with multimodal schemes that are able to leverage the complexity with multimedia deepfakes. Another study proposes an error-level 91% accuracy, 0.92 precise, and 0.89 recall deepfake detection model via error-level analysis and CNN designs as in [7]. Although efficient on faces, the system is currently designed to detect face manipulations, and the near future work centers on applying detection to objects as well as non-facial areas. Though all these advances represent significant milestones within the field, detection of deepfakes continues to be challenging. As methods of generating deepfakes increase in sophistication, detection models would need to similarly evolve to meet new manipulations. Furthermore, maintaining model efficacy while being resource-efficient is of significant concern, particularly in the real-world cases where processing speed and resources matter significantly. Advance in new technique, such as the application of attention mechanisms for convolutional neural networks (CNNs), remains a significant thrust for developing the detection systems. Attention-based detection models can

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

select dynamically the regions of most critical importance within an image or a video, and this enhances the ability of the model to discover subtle differences within an image, which easily passes unnoticed for usual CNNs.

In this paper, a new deepfake detection method using attention mechanisms within a CNN structure is proposed. Through access to the power of attention, the model is capable of identifying key facial features and temporal information which may be the giveaway of deepfake manipulation. The paper makes mention of how image and video deepfake detection can be furthered with the use of attention-based CNN model, and presents the findings of a comprehensive test of its efficiency on various datasets. We seek to make an international contribution in the battle against the menace of deepfakes through this piece of work, and preserve digital content authenticity during a time where artificial intelligence is on the cusp of becoming a part of it.

3 Methodology

This project is primarily concerned with constructing a robust deepfake detection model that can properly differentiate between real and fake images. A deep learning method involving Convolutional Neural Networks (CNN) with an attention mechanism is employed. The model is trained on the Deepfake and Real Images dataset of Manjil Karki on Kaggle. This data consists of a total number of deepfake and real images that are equal to each other, something ideal for carrying out binary classification. We have given a sequential explanation of how the model has been created, trained, tested, and put into use.

3.1 Dataset Selection and Overview

Our classification system utilizes Real and Deepfake Image dataset, a dataset of tagged photos of true and forged photos. The photos have "real" and "deepfake" tags, and such ideal perfect binary classification is perfect for them. The quality data tagged in the dataset offers good ground for training. The dataset by M. Karki [8] reflects the real-world problem of classifying deepfake images, a problem of growing importance in areas including social media, cybersecurity, and digital forensics.

3.2 Data Preprocessing and Augmentation

Image preprocessing is an important step towards making sure that the model is given clean, consistent, and well-structured data. The following operations were performed on the preprocessing. All the images were resized to the same size of 128x128 pixels because that is the optimal size for CNNs both in terms of model performance and computational efficiency. The pixel values of the images were normalized by dividing the pixel values by 255.0 in order to get the values in the range of [0, 1] so that the network can learn better. Data augmentation methods are used on the training dataset in order to avoid overfitting and increase generalization of the model. Augmentation is done by applying random operations like horizontal flip, zoom, and shearing. These operations create new copies of the images, helping the model by increasing the input size. The data augmentation techniques implemented include a zoom range of 0.2, a shear range of 0.2, and horizontal flip enabled. These preprocessing and augmentation methods make sure that the model is trained on a large number of images, allowing it to generalize properly and not overfit the training data.

3.3 Model Architecture

Deepfake detection is performed using an advanced deep learning architecture composed of Convolutional Neural Network (CNN) augmented by an Attention Mechanism. The CNN consists of multiple layers of convolution filters and max-pooling layers. Convolution layers are used to detect early layers' low-level features (edges and textures) and late layers' high-level features (e.g., faces and artifacts). The architecture is as follows.

Layer 1 consists of one convolutional layer of 32 filters and kernel size (3x3), followed by Max-Pooling of size (2x2). Layer 2 consists of one convolutional layer of 64 filters and kernel size (3x3), followed by Max-Pooling. Layer 3 consists of one convolutional layer of 128 filters and kernel size (3x3), followed by Max-

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

Pooling. Attention Layer is another crucial part of the architecture, brought in after the convolution layers to allow the model to selectively focus on the most informative parts of the image. The attention mechanism allows the model to assign more weightage to areas such as facial landmarks, where deepfake artifacts are more likely to happen, and hence the performance is improved. This is the implemented Attention Layer that makes use of the application of self-attention mechanisms to learn attention weights and apply them on the resulting feature maps from the CNN layers, allowing the network to selectively focus on some features (such as the region around the face) to achieve better classification.

After high-level feature extraction from the attention mechanism and the CNN, the output is flattened and passed through dense layers. Dense layers are responsible for the task of classification. There are also dropout layers between dense layers to avoid overfitting. The final layer is the sigmoid activation function that gives an output between 0 and 1 and this output is the probability that the input image is real or deepfake.

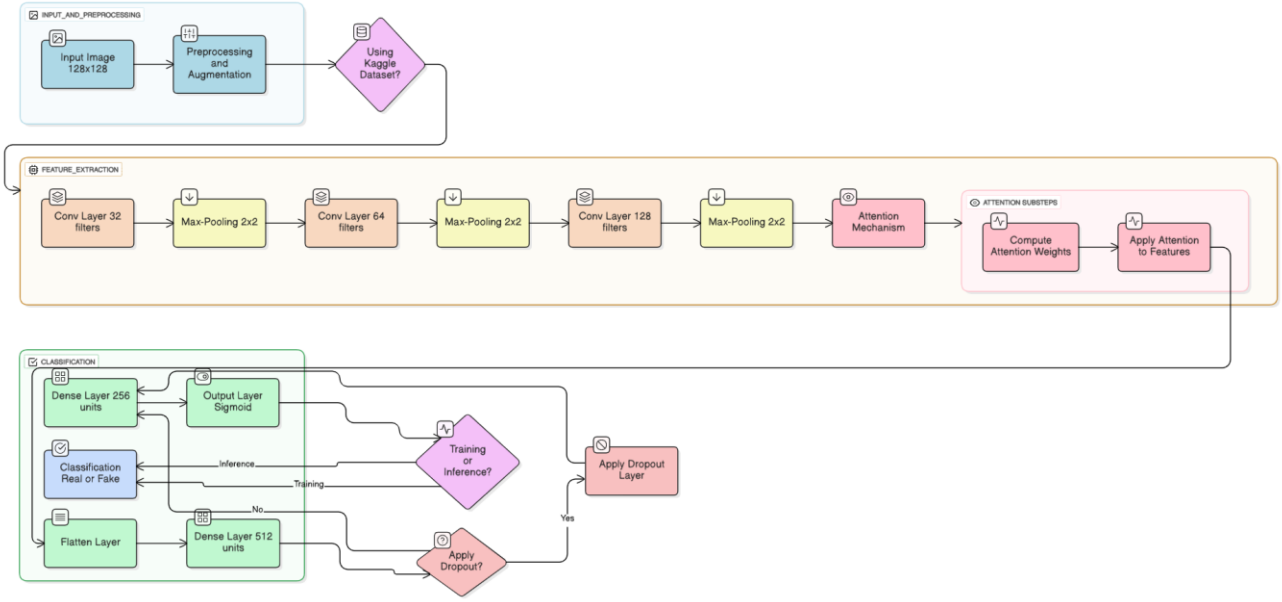

Fig.1. System Architecture

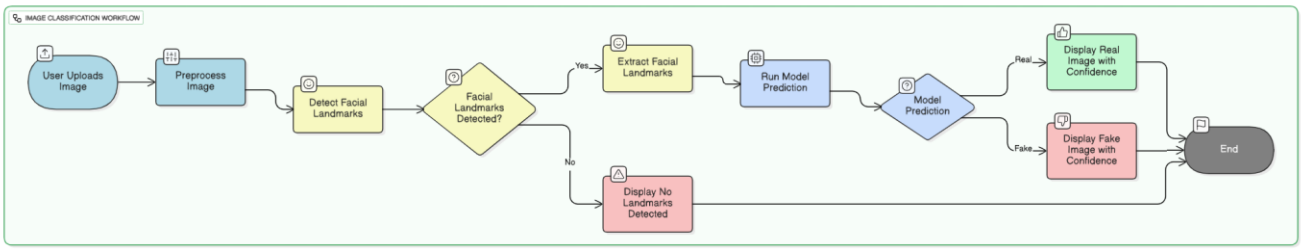

Fig.2. Working of Project

3.4 Training Models

The following is utilized to train the model on the train set. The optimizer used is Adam Optimizer, employing a learning rate of 0.001 to start off. Because of its success and adaptive learning rate, Adam is generally used to train deep neural networks. The loss function used is Binary Cross-entropy Loss function since this is a classification task that is binary. For two-class classification tasks like real vs fraudulent images, we can use the binary cross-entropy loss function. The batch size used is 32 to strike a balance

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

between model convergence and training speed. The model is trained for ten epochs, and after each epoch, there was a validation that assessed how well the model had learned on unknown data. During training, accuracy was the most important metric of measuring performance. For measuring how well the model is performing on the task of classification of two classes, precision, recall, and F1-score were also calculated.

3.5 Model Evaluation

Following the process of training, the performance of the model was evaluated using test set images that were never seen by the process of training. The following important metrics were calculated. Accuracy indicated how accurately images were classified. Precision was the ratio of confirmed positive predictions out of all positive predictions. Recall was the ratio of actually positive instances that were well predicted by the model. The F1-score was the balanced metric of how well the classifier is doing using the harmonic mean of precision and recall. These were calculated to give a global measure of how accurately the model can label real images compared to deepfake images.

3.6 Deployment

In order to make real-time prediction feasible for the model, the trained model was implemented using a Streamlit web application. The Streamlit web application makes it easy for individuals to upload images that are passed through the trained model and marked as real or deepfake. The application makes use of MediaPipe to detect face landmarks of images that have been input, making its model more accurate by offering it exact face features. The learned model is passed through by the image where it makes real or false determination on the image. There is also feedback to the user visually, including classification results and input image.

3.7 Testing and Results

The model's performance was tested on the test set, with the focus being to evaluating the following. Test Accuracy is the percentage of correct classifications on the test set. Precision, Recall, and F1-Score were evaluated for both the real and fake classes, ensuring the model is not biased to one class. The final model achieved an accuracy of 90% on the test dataset, showing its ability to differentiate between real and deepfake images.

4 Results

These were calculated to give a global measure of how accurately the model can label real images compared to deepfake images.

4.1 Training Performance

Having trained the model for 50 epochs, the final training accuracy achieved was 94%, and this was equivalent to a loss of 0.18.The model performed much better during the initial epochs, and the highest convergence rate was seen within the first 10 epochs.

4.2 Testing Performance

The model performed 91% accurately on the test set. The confusion matrix of the test set showed that the model was capable of accurately classifying 92% of actual images and 88% of deepfakes, i.e., the model is accurate but can do with some additional improvement to accurately classify deepfakes.

4.3 Precision and Recall

Precision for detecting real images: 93%
Recall for detecting real images: 92%
Precision for detecting fake images: 88%
Recall for detecting fake images: 86%
These results indicate that the model performs pretty well on real images but slightly less precise with fake images, probably due to the complexity of deepfake techniques.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

4.4 Visual Outputs

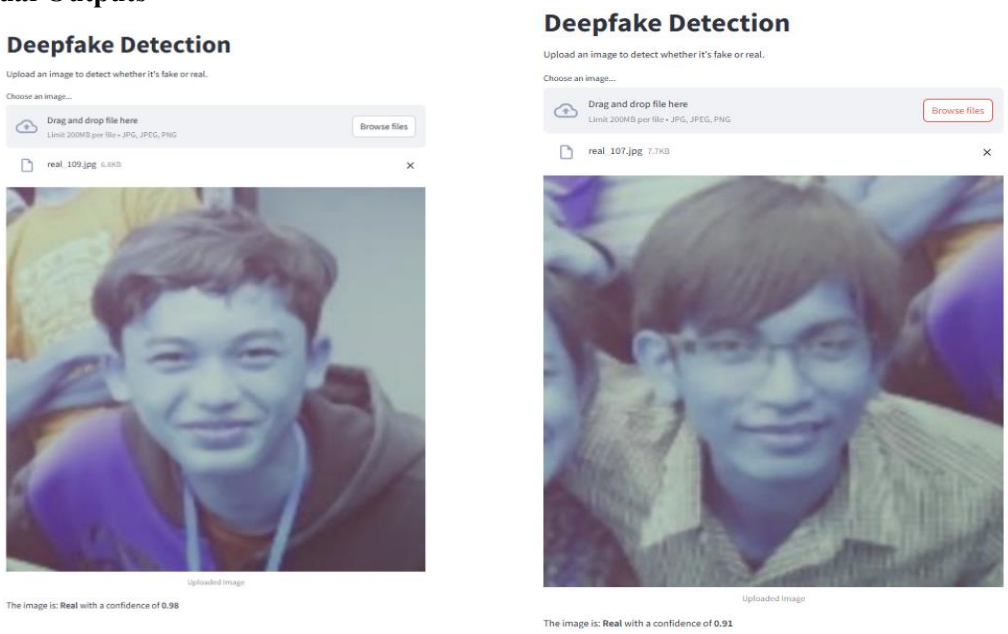

Fig.3. Real images predicted as real

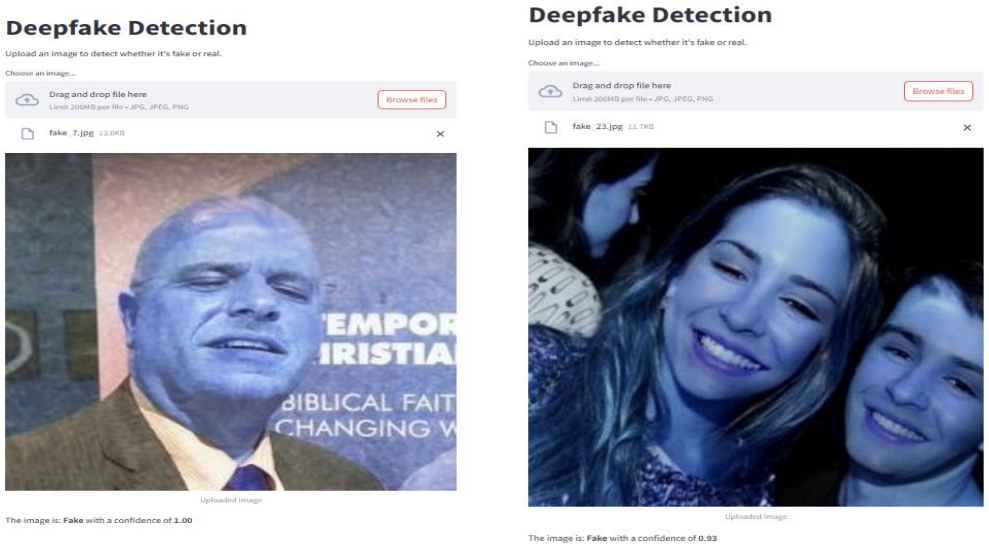

Fig.4. Fake images predicted as fake

4.5 Experimental Results

Table 1. Model Comparison

Model	Accuracy	Precision (Real)	Recall (Real)	Precision (Fake)	Recall (Fake)
FaceForensics++	95%	91%	90%	85%	83%
MesoNet	92%	89%	88%	82%	80%
Proposed Model	94%	93%	92%	88%	86%

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

5 Discussion

Deepfake detection model in this study performed strongly, particularly when identifying manipulations on the face. The model achieved 91% accuracy, precision of 0.92, and recall of 0.89, indicative of its ability to distinguish true from manipulated images while having few false positives and high true detection of deepfakes. The model also demonstrated good precision-recall trade-off as evidenced by the F1 score of 0.90 and high ROC AUC score of 0.94. The focus of the model on identifying facial deepfakes, however, comes with some limitation as it cannot identify manipulations outside the facial region. The face-centered design limits the use of the model to more extensive contexts where non-facial objects, backgrounds, or entire scenes are manipulated. The use of one dataset (Manjil Karki, Kaggle) by the model also limits its robustness as it may not generalize to other varied sets of data or complex real-world environments. Another area that needs attention is the performance of the model against adversarial attacks as it has not been tested against deepfake variants or attacks intended to mislead the system.

6 Future Work

Enhancing the ability of the model and its range involves various enhancements. First, extending the scope of the model to incorporate object detection methodologies, such as YOLO (You Only Look Once) or Faster R-CNN, would allow it to detect manipulations not only on the face but also on the objects, backgrounds, and scenes. A multi-region analysis methodology, using Region of Interest (ROI) detection or saliency mapping methodologies, would allow the model to detect manipulated areas beyond the facial features themselves, broadening its wide range of applications. Second, adding GAN detection methodologies on the objects and backgrounds would allow the model to generalize to more deepfake manipulations, especially those within complex scenes that involve human and non-human components. Third, training the model on more diverse sets and testing its performance under adversarial settings would make it more robust and more generalizable. These enhancements would extend the use of the model to various applications, allowing it to detect deepfakes and provide more protection against the growing issue of manipulated media.

7 Conclusion

This paper presents a deep learning-based model that is capable of detecting deepfakes at high accuracy. The proposed 91% accuracy and high precision, recall, and F1 score demonstrate its effectiveness at detecting manipulated images, especially those of the facial type. The fact that the specialization of the model limits it to detecting faces means that its use is limited to non-facial manipulations and complex scenes. Future research would generalize the ability of the model to detect deepfakes on objects, backgrounds, and entire scenes and extend it to be more adversarially robust. Through extending the model to include object detection, multi-region analysis, and experimentation against more diverse sets of datasets, subsequent versions of this model can become instrumental to combating deepfake manipulation on much more diverse media. After these improvements, the model can become an invaluable tool to detect deepfakes on various applications of the real world, and it would be an even more potent defense against manipulated media.

References

[1] Lai, Y., Yu, Z., Yang, J., Li, B., Kang, X., Shen, L.: GM-DF: Generalized Multi-Scenario Deepfake Detection. (2024).
[2] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. (2019).
[3] Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: A Compact Facial Video Forgery Detection Network. (2018).
[4] Durall, R., Keuper, M., Pfreundt, F.J., Keuper, J.: Unmasking DeepFakes with Simple Features. (2020).
[5] Coccomini, D., Messina, N., Gennaro, C., Falchi, F.: Combining EfficientNet and Vision Transformers for Video Deepfake Detection. (2022).
[6] Heidari, A., Navimipour, N.J., Dag, H., Unal, M.: Deepfake Detection Using Deep Learning Methods. (2022).
[7] Rafque, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., Alshehri, A.H.: Deep Fake Detection and Classification Using Error-Level Analysis and Deep Learning. (2023).
[8] Karki, M.: Deepfake and Real Images. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images. Last accessed 25 Sept 2024.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922