Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

DeFakeNet: A ResNet50V2-Based Deep Learning Model for Deepfake Detection and Classification

Debasish Samal[1], Dimple Nagpal[2], Prateek Agrawal[3], Vishu Madaan[4], Wou Onn Choo[5]

[1]Research Scholar, School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India.
[2]Assistant Professor, School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.
[3]Professor and Dean Academics, SGT University, Gurugram-Badli Road, Chandu, Budhera, Gurugram, Haryana, India.
[3]Research Fellow, INTI International University, Nilai, Malaysia.
[4]Associate Professor, School of Engineering and Technology, SGT University, Gurugram, Haryana, India.
[5]Associate Professor, Faculty of Data Science and Informative Technology, INTI International University, Nilai, Malaysia.

E-mail: [1]debasishsamal01@gmail.com, [2]dimplenagpal009@gmail.com, [3]dr.agrawal.prateek@gmail.com, [4]dr.vishumadaan@gmail.com, [5]wouonn.choo@newinti.edu.my

Orcid ID: [1]0000-0002-0217-5221, [2]0000-0002-6787-9078, [3]0000-0001-6861-0698, [4]0000-0002-9127-4490, [5]0000-0001-8397-3251

Abstract

The spread of deepfake content has created distrust, misconception, and fraud all around the world. Swapping faces between individuals seamlessly or generating AI-created fake faces has become easier than ever with AI tools like DALL-E, Midjourney, ChatGPT, and Google Gemini. AI generated obscene and malicious content has become progressively prevalent and widely circulated on social media due to the misuse of generative adversarial techniques. To prevent AI generated fake images from causing harm to the identity and social integrity of a person or community, this research presents a deep learning model called 'DeFakeNet' based on the advanced ResNet50V2 CNN architecture, designed to detect and classify whether a person's face is real or fake. While past research has relied extensively on pre-trained models and limited dataset, DeFakeNet was trained on a custom-developed dataset titled 'Real vs Fake Faces Balanced Dataset with Multiple Dataset Splits', a mixed dataset comprising 10,000 high quality balanced real and fake face images. Upon testing with unseen data, the proposed model obtained 91.95% test accuracy and an AUC score of 97.64%, setting new records in this field. A critical scrutiny of all the diverse evaluation metrics, ROC and Precision-Recall Curves is presented in this paper, which is rarely discussed thoroughly in previous research. Additionally, the model performance comparison with current methods shows robust real world reliability and application toward the detection of evolving deepfakes.

Keywords: DeFakeNet, Deep Learning, Deepfake Detection, Peaceful Society, Social Security, Social Safety.

DOI: https://doi.org/10.36548/jiip.2025.4.015
Received: 01.10.2025, received in revised form: 27.10.2025, accepted: 10.11.2025, published: 21.11.2025

1. Introduction

Deepfake technology has rapidly progressed to creating highly realistic multimedia of individuals fabricating comments or behaviours. This technology is linked with substantial security and privacy concerns. For instance, a deepfake video could be created that makes it seem like a celebrity is promoting a brand or making offensive statements [1]. Such a situation creates substantial concerns regarding a person's personal and professional reputation. Deepfakes can imitate one's facial expressions, talking style and voice to deceive others into spreading private information. At the pace at which the technology is advancing, it is evident that certain face forensics protocols or reliable systems should be invented to tackle this global issue [2].

Figure 1. Deepfake Face Instance, (Left Image) Real And (Right Image) Fake using FaceApp.[Source Webpage: https://www.faceapp.com]

Figure 1 shows that using FaceApp, one can change the face of a person's real image and its younger artificial counterparts.

To counter this threat, deep learning techniques are crucial for effective deepfake detection and classification [3]. Notably, Convolutional Neural Networks (CNNs) are the cutting edge approach for image based detection and classification tasks as these architectures are inherently designed to analyse visual data. The layers of a CNN model learn hierarchical spatial features. CNNs are known for identifying patterns ranging from simple textures and edges to complex shapes automatically. In the context of deepfake detection, these convolutional layers are able to learn the subtle, low-level artifacts, inconsistent patterns and unnatural skin textures which are traits of AI-generated fakes, often indiscernible to human eye. These are widely used to compete against AI techniques like Generative Adversarial Network (GANs) [4]. The classifications should be accurate and versatile to detect various kinds of manipulation techniques, especially GAN generated fake images [5]. Identifying potential risks in advance and securing a person's identity is essential to protect people vulnerable to fraud and constant cyberattacks. The development of trustworthy detection technologies is. Therefore, essential to address these risks and enhance digital security [6].

1.1 Motivation

The following research is motivated by the massive amount of false multimedia circulating online, undermining the social safety and security of anyone prone to be blackmailed or threatened with death without committing any crime. Innocent people are being harassed with deepfake versions of themselves, with images showing things with false agendas. The risks are far beyond our imagination with the widespread use of deepfake technology in pornography, individuals may find forged images or videos depicting them in explicit acts they have never been a part of. The protection of one's privacy, especially identity protection is increasingly critical in an era where visual inspection often dictates belief and perception. To counter this misuse of deepfake technology, this research proposes a CNN based model which is reliable for accurately detecting realistic looking fake face images.

Because of its proven capacity to successfully address image-based binary classification problems, the suggested CNN architecture is especially well-suited for this purpose. Its optimal size has no effect on its exceptional accuracy in real-time computer vision applications, such as image classification. It works well for spotting deepfakes in a variety of datasets. A reliable deepfake detection system will need to evolve as deepfake creation techniques advance.

1.2 Main Contribution

The research can be summarized with the following contributions:

- Establishment of a novel deep learning model, 'DeFakeNet', specially developed for deepfake image detection, taking advantage of ResNet50V2 CNN model architecture to obtain high accuracy.
- The core novelty includes experimenting with a custom-developed dataset that has been systematically restructured from the existing RVF10K dataset into balanced multiple splits which enhances model stabilization and addresses the need for an optimally structured dataset.
- A practical validation of the model's real-world credibility through a detailed inference time analysis (approx. 9.3ms per image), confirming the feasibility for deployment in real-time detection systems.
- This research performs a thorough review of all crucial evaluation criteria, in contrast to normal evaluation procedures found in earlier studies. A comprehensive and open assessment of the model's reliability is ensured by the detailed discussions and presentations of the Confusion Matrix, Matthews Correlation Coefficient (MCC), and the entire range of ROC and Precision-Recall Curves.
- Visual binary image classification through model prediction and comparative assessments against current methods confirms DeFakeNet's robust effectiveness and its suitability for real-world applications against evolving generative AI threats.

The detailed outline for the paper contains a brief discussion of existing deepfake detection methods in Section 'Literature Review'. The research provides methodology in Section 'Proposed Work'. Furthermore, the outcomes of assessing the proposed model are described in Section 'Results and Discussion'. It also contains a comprehensive analysis of evaluation metrics and a comparison of DeFakeNet's performance with existing research

works. In Section 'Conclusion', the proposed investigation delivers a summary of the research and suggests possible areas of investigation for subsequent studies in the future.

2. Literature Review

In recent years, the identification of manipulated media, specifically AI-generated synthetic images and videos has been executed by recognizing and classifying deepfakes using machine and deep learning techniques. Traditional techniques allowed machine learning models to process manually created characteristics from still images or video sequences [7-9].

Deep convolutional neural networks (DCNN) that rely on images are widely used for detecting deepfakes and performing image classification tasks [10]. In the meantime, the CNN model can identify anomalies like inconsistent expressions or unusual textures resulting from the deepfake process. This can be achieved by expanding CNN to capture temporal irregularities between multiple frames in video deepfake cases.

Pre-trained autoencoders can learn features from a set of images and compare them with a reconstructed image generated by a trained model [11]. The reconstructed image may have been modified if there are any inconsistencies between it and the original. It is specifically utilised to detect low-level artifacts discovered in deepfakes.

Goodfellow et al. [12] employed GANs to simultaneously generate and detect deepfakes. To enhance overall performance and improve generalization when identifying adversarial content, these methods utilize a discriminator (the "D" in the GAN architecture) trained to differentiate between authentic and generated images.

Recently, Vision Transformers (ViTs) have surfaced as a powerful alternative to CNNs, adapting the transformer architecture, originally from natural language processing (NLP), for image analysis [13]. ViTs can more easily detect minor artifacts or failures due to their ability to embed an image's global information. In the context of deepfake photos, the use of pre-trained models on deepfake datasets combined with transfer learning, is an especially useful feature.

Petmezas et al. [14] implemented a hybrid model and combined the predictions of various models, including CNNs, RNNs, and ViTs, to enhance detection accuracy. Ensemble methods utilise a range of techniques [15], thereby combining the distinct benefits of each to form a more comprehensive defense against deepfakes, especially when a broad variety of manipulations are employed. The effectiveness of these techniques varies depending on the type of deepfake and the context.

Table 1. Summary of Recent Deepfake Detection Related Research Works

Reference	Method Summary	Advantage	Weakness
[9]	Manually Extracting features: Detects irregularities in movements, face characteristics and facial features.	Highlighting distinct visual discrepancies is the most effective for specific types of manual alterations.	The system faces constraints in its ability to expand, and it is also ineffective against AI-generated deepfakes.

[10]	CNNs based on Images: It uses convolutional layers to identify anomalies like inconsistent expressions or textures.	The model shows strong performance in image classification tasks and detects spatial anomalies in both images and video footage.	Temporal information is typically not captured automatically and requires significant datasets for effective training.
[11]	Autoencoder-decoder based detection: The comparison of original and reconstructed images is used to detect discrepancies	Focused on detecting minor discrepancies, it centres around errors in the reconstruction method.	The performance of the autoencoder severely relies on the quality of the model used and the dataset.
[12]	Use of GAN discriminators to distinguish between authentic and artificially created content.	Training with adversarial examples improves model robustness. Generalizes effectively to various types of content.	This kind of system continues to be resistant to a broad spectrum of manipulations. Adversarial training necessitates substantial adjustment of hyperparameters.
[13]	ViTs encodes global image information and applies transfer learning to pre-trained deepfake datasets.	This method excels at global representation and transfer learning capabilities, proving to be particularly effective at identifying minor artifacts.	ViTs require computationally intensive processing and adjustments for particular deepfake data sets.
[14,15]	Hybrid model combines predictions of CNNs, RNNs, and vision transformers.	Increases precision by unifying the key strengths of multiple models and is resilient to various forms of manipulation.	The rising complexity and computational demands of models require comprehensive integration and high computational training cost.

The literature review, summarized in Table 1 highlights several studied approaches, summarizing their methods, benefits, and drawbacks. Despite the impressive progress achieved by the existing methods, several limitations remain. Complex models like ViTs and ensemble methods show high accuracy but require high computational and training costs, which makes them difficult to run in real-world scenarios or on resource constrained devices. On the other hand, traditional CNNs, pre-trained networks and autoencoders are computationally feasible but highly dependent on the scale and quality of the training data. Using lightweight CNN models may struggle to generalise against evolving generative techniques. Much of the prior research is restricted by a dependence on limited dataset which prohibits robustness and generalization, combined with a lack of multi-metric comprehensive evaluations that make existing solutions difficult to assess. To address these critical issues, this research introduces

'DeFakeNet' which leverages the residual learning mechanism of ResNet50V2, with training and testing conducted on a high quality optimal dataset for effective deepfake detection and classification.

3. Proposed Work

This section explains the proposed DeFakeNet, CNN architecture, which is built on the advanced ResNet50V2 model, as illustrated in Table 2. In contrast to more complex designs, this model reduces overfitting on relatively smaller datasets and is computationally feasible for environments with limited resources, such as devices with modest GPU capabilities. The proposed model in total possesses 25,664,001 parameters. The flowchart diagram of the proposed methodology contains two steps, model training and testing as depicted in Figure 2.

The ResNet50V2 architecture was particularly chosen as the backbone of DeFakeNet due to its balance of deep feature extraction capabilities and computational feasibility. Older, more complex architectures like VGG16 may suffer from vanishing gradients in deep networks, The residual skip connections feature of ResNet50V2 allows for efficient gradient flow, enabling the training of much deeper networks. This critical choice enables the model to detect subtle low-level artifacts and textural inconsistencies. Compared to lightweight models like MobileNet variants or the EfficientNet family of architectures, the framework of ResNet50V2 provides a resilient feature hierarchy offering a powerful baseline for transfer learning in binary classification.

The following subsections describe image data collection and preprocessing, which incorporates techniques such as image augmentation, model design, training configurations and hyperparameter tuning to achieve optimal performance and resilience in differentiating real and deepfake images.

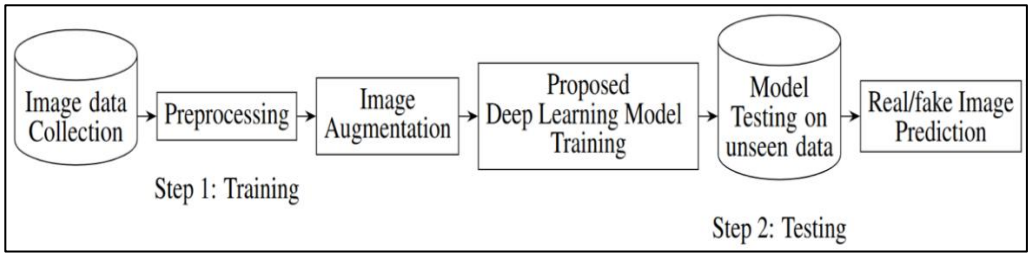

Figure 2. Flow of Proposed Methodology Diagram for DeFakeNet Model Training and Testing

3.1 Image Data Collection

'Real vs Fake Faces Balanced Dataset with Multiple Dataset Splits' is a diverse mixed dataset containing 10,000 high quality face images. There are mainly four balanced training and testing dataset splits (60-40, 70-30, 75-25, 80-20) that make up the whole dataset. The chosen dataset is a restructured and enhanced version of the existing RVF10K dataset from Kaggle [19]. Initially, the RVF10K dataset contained only two subdirectories (train, valid) and was originally in a 70-30 data split. Later, the dataset was restructured into balanced data split ratios to support research experiments, and its updated form has been uploaded and is freely accessible on Zenodo [20] to enhance reproducibility and foster further research in deepfake

detection. The restructuring process was executed using the stratified sampling method which ensures that a 50:50 class balance is preserved in every split of the dataset. No over or under-sampling techniques were used to structure the balanced data splits. Subject independence was enforced so that no individual's images appeared in more than one set, thereby preventing data leakage.

For this research, the 60-40 ratio of the dataset split has been utilized. The rationale for using this specific split was to achieve a balance between model robustness during training and the reliability of the evaluation results. By training the proposed model with 60% of the data, it allows the model to be exposed to a diverse range of real and fake facial characteristics, which in turn facilitates effective feature extraction. The remaining 40% of the dataset is utilized as a comprehensive and unbiased test set to assess model performance and prevent overfitting of the training data. The labels of the images for the dataset were inherited from the original RVF10K dataset, which included real face images from the Flickr-Faces-HQ (FFHQ) dataset and fake face images generated from StyleGan. The restructured mixed dataset preserved these original, validated labels. A summary of the used dataset division is displayed below in Table 2.

Table 2. Dataset Split Images Used for Deepfake Detection Experiments in this Research

Dataset Split (60-40)	Real Images	Fake Images	Total Images
Training Set	3,000	3,000	6,000
Validation Set	1,000	1,000	2,000
Testing Set	1,000	1,000	2,000
Total	5,000	5,000	10,000

Preprocessing: The two classes, real and fake images are pre-processed to standardize the input dimensions and normalize the pixel values. All are scaled to a higher pixel size of 224x224 to maximize the effectiveness of the input model feature extractor for the experiment. The specific 224x224 input resolution was selected because it is the standard input dimension in which the ResNet50V2 model was pre-trained on the ImageNet dataset. Utilizing the original input size is a core requirement of transfer learning, as it allows the model to correctly process features using the learned weights. Normalization of images occurred by modifying pixel values to a range of [0,1]. This enables faster convergence during training by keeping them within a standard range.

Image Augmentations: To amplify the versatility of images and training performance, data augmentation is used which includes up to 10 degrees of random rotations, 20% width and height shift, flipping images horizontally and vertically, 20% zooming, and shearing. These techniques help the model generalize across diverse inputs and reduce any chances of overfitting occurring.

Table 3. DeFakeNet Deep Learning Model Summary with Hyperparameter Tuning

DeFakeNet Model Layer (type)	Output Shape	Number of Parameters
InputLayer	(None, 224,224, 3)	0
ResNet50v2	(None, 7, 7, 2048)	23,564,800

Global_average_pooling2d (GlobalAveragePooling2D)	(None, 2048)	0
Dense	(None, 1024)	2,098,176
Dense_1	(None, 1)	1,025
Total Parameters:		25,664,001
Total number of layers:		195
Trainable:		25,618,561
Non-trainable:		45,440
Hyperparameters		Values
Input Image size		224x224
Optimizer		Adam
LR (Learning Rate)		0.0001
Batch Size		32
Total Epochs		40
Loss		binary_crossentropy
Activation		Sigmoid

3.2 DeFakeNet Model

The DeFakeNet model was designed on the ResNet50V2 architecture for its deep residual learning framework. The proposed deepfake detection model uses transfer learning to increase its detection and classification accuracy. As displayed in Table 3, the model processes input images, which are 224x224 pixels in size, with a colour depth of 3.

To adapt the pre-trained ResNet50V2 model to the specific task of deepfake detection, the entire base model was unfrozen and allowed to be fine-tuned for training as reflected in the 25.6 million trainable parameters shown in Table 3. This fine-tuning process adjusts the model's learned features to make it highly specialized in identifying deepfake artifacts. The 45,440 non-trainable parameters relate to the batch normalization layers, which were kept frozen to maintain stable feature distributions.

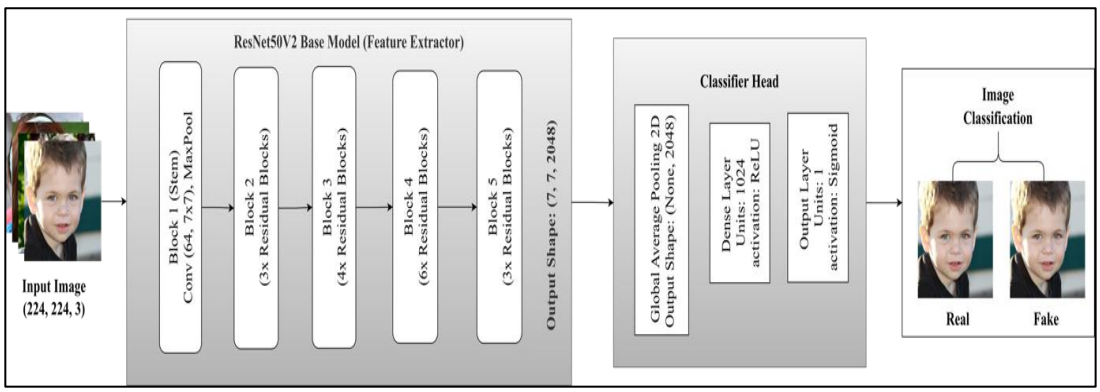

Figure 3. DeFakeNet Model Architecture

The feature maps extracted by the ResNet50V2 base model are processed using a Global Average Pooling (GAP) layer. This layer is an essential component of the model as it significantly reduces the number of trainable parameters compared to a traditional 'Flatten' layer. By averaging the values of each feature map into a single number, this layer generates a fixed 2048 size vector regardless of the input's spatial dimensions. This structural choice prevents overfitting and enables the model classifications to be more robust to spatial translations of features within the face image.

A subsequent dense layer of 1024 neurons with The ReLU activation function was used to facilitate sufficient capacity for the model to learn complex non-linear combinations of the features extracted by the GAP layer. This dense layer size balances model complexity with the potential risks of overfitting, which can occur with overly large dense layers.

3.3 Model Training

This training with model tuning adjustments contributes to stabilizing the training process and results in enhanced convergence for the model. The process uses a batch size of 32, which enables efficient use of computational resources and balances memory limitations. The binary crossentropy is employed for training as loss function, suited for binary image classification tasks.

A summary of the hyperparameters utilised and their respective values is presented in Table 3 along with the model summary. The training was conducted for a maximum of 40 epochs, incorporating 'Early Stopping' callbacks. It involves monitoring validation loss and training stops if no progress occurs within 5 consecutive epochs. This helps prevent overfitting and maintains the best possible weights. Learning rate scheduling was implemented so that if the validation loss does not show improvement over 3 epochs, it automatically reduces the learning rate by a factor of 0.5, down to a minimum of 1×10^{-6}.

3.4 Experimental Configuration

The training, testing and evaluation are performed on a Windows 11 OS with an Intel Core i7 powered CPU and GeForce RTX 3060 GPU by NVIDIA.

3.5 Evaluation Metrics

DeFakeNet performance is measured across multiple evaluation metrics, including accuracy, recall, F1 score, precision, true positive rate (TPR) and false positive rate (FPR).

The accuracy (eq.1) is calculated as:

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

where TP, TN, FP and FN refer to the total count of true positives, true negatives, false positives, and false negatives.

Precision (eq.2) is specified as:

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

Recall (eq.3) is calculated as:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \tag{3}$$

F1-Score (eq.4) is formulated as:

$$\text{F1 Score} = \frac{2 \text{ x } (Precision \ x \ Recall)}{(Precision + Recall)} \tag{4}$$

Classification Error is measured as:

$$\text{Classification Error} = \frac{(\text{FP+FN})}{(\text{TP + TN +FP + FN})} = 1 - Accuracy \tag{5}$$

The classification error is the proportion of all incorrect predictions to the total number of predictions. It can also be the complement of the classification accuracy as shown in eq. (5).

4. Results and Discussion

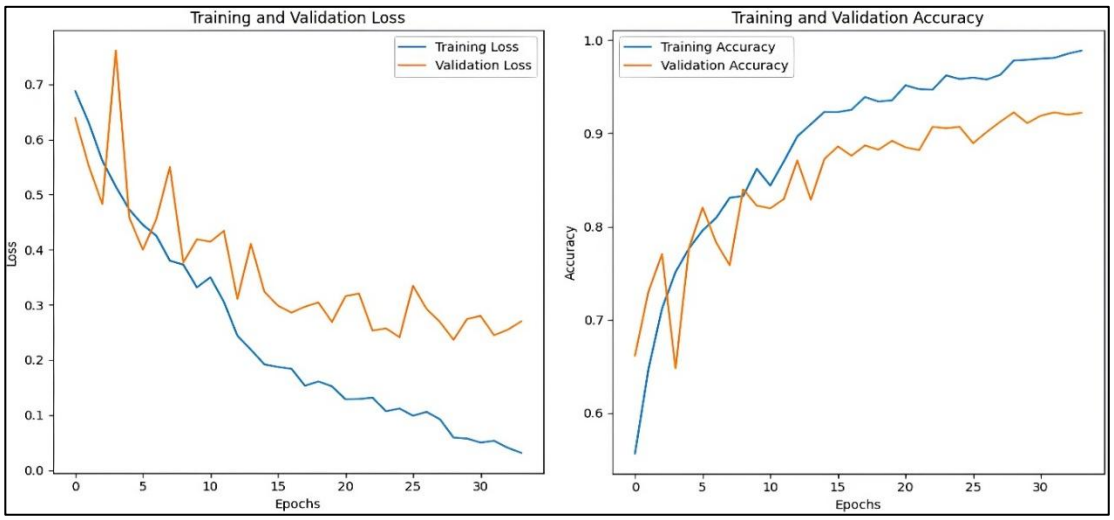

Figure 4. Training and Validation Loss graph (Left), Accuracy Graph (Right) of 'DeFakeNet'

The training was run for approximately 40 epochs. In the initial epochs, the model showed quick convergence with high classification accuracy. Figure 4 represents the graph of the proposed model accuracy and loss during training and validation. By the end of epoch 34, a high accuracy of 98.88% was achieved by DeFakeNet model, while the validation accuracy also crossed 90% and achieved 92.20%. The gap between the high training accuracy and the final validation accuracy indicates that a degree of overfitting, which is a common difficulty when a deep network learns the features of the training image set. However, the utilization of data augmentation, learning rate scheduling and early callbacks was implemented specially to tackle this. The early stopping with a patience of 5 epochs, effectively halted the training at epoch 34 after achieving the lowest validation loss of 0.23. This assured that the saved model weights were from the epoch with the best performance rather than the epoch with the highest training accuracy. As a result, the process prevented a significant degradation of the model's ability to generalize to new input face images.

The observed fluctuations in the validation loss as seen in Figure 4 (left), convey that the model is encountering more challenging examples in the validation batches. While the

overall trend is decreasing, this limited rate of instability highlights the difficulty of the classification task. To strategically manage this behaviour and ensure a smooth training process, a dynamic learning rate scheduler and early stopping were implemented. In the beginning of the model training, the learning rate was set to 1.0 x 10-4. For the first 12 epochs the learning rate remained unchanged while the training accuracy increased from 55.65% to 86.98% and simultaneously, the validation accuracy improved to 82.95%.

The learning rate was subsequently decreased to 5.00 x10-5 during epochs 13-28 and then reduced further to 2.50x10-5 and 1.25x10-5 in epochs 29-34, resulting in refined weight adjustments and additional performance improvements as well as enhanced stability. At epoch 29, the lowest validation loss of 0.2365 was found, which corresponded with one of the peak validation accuracy values of 92.25% at epoch 32.

This research analyses the performance on the hardware specified in subsection 3.4 and found that the complete training process took approximately 33.2 minutes (1991 seconds). The time per epoch was very consistent, averaging about 55-60 seconds after the first one (77s). This efficient training time demonstrates that the DeFakeNet model can be rapidly retrained. For practical deployment of the model, the essential metric is inference time. During training, the final epoch showed an average processing speed of 297 milliseconds(ms) per step (batch) where each batch consists of 32 images. This corresponds to an average inference time of approximately 9.3ms per image. This high-speed inference performance, capable of processing over 100 images per second (1000ms / 9.3ms = 107.5), validates that the DeFakeNet model is highly suitable for real-world application and can be feasibly deployed in real-time systems such as scanning images upon uploading to social media or inclusion into rapid digital forensic pipelines, where both speed and accuracy are crucial. Collectively, these results show that DeFakeNet can effectively identify real and fake face images with minimal evidence of overfitting, thereby positioning it as a highly competitive approach for this task.

Confusion Matrix Analysis:The confusion matrix compares the predictions made by the proposed model with the testing set containing a balanced mix of a total of 2,000 face images. As shown in Figure 5, out of the 1,000 fake face images, DeFakeNet correctly identified 937 images as fake, and 63 images were misclassified as real images. Similarly, from the 1,000 real images, 902 images were accurately identified as real, and only 98 images were wrongly classified as fake when they were actually real.

From the confusion matrix, the evaluation metrics were calculated. For this analysis, the 'Real' class was treated as Positive and the 'Fake' class as Negative. The resulting Precision was 93.47%, which means that of all images predicted as 'Real', 93.47% were correct. The Recall score was 90.20%, indicating that the model correctly identified 90.20% of all real images. To specifically address the fake detection rate, the research study calculated the True Negative Rate (Specificity), i.e., the recall of the 'Fake' class. This was calculated as

(TP_fake / (TP_fake + FN_fake)), which is 93.7%. The F1-Score was 91.81. This meticulous analysis also reveals a slight bias in the model predictions. The class imbalance indicates a slight tendency to misclassify an image as 'fake' due to the misclassification of 98 'real' images as 'fake' and only 63 'fake' images as 'real'. To handle the class imbalance, the research study implemented balanced metrics like F1-Score and ROC AUC curves. This trade-off, while improving the detection of fake images, comes at the cost of a slightly higher rate of false alarms on real images. While the test accuracy was 91.95%, the classification error rate was 8.05% (1 - 0.9195), meaning the proposed model incorrectly classified only 8.05% of the

total test set. Overall, the model's classification results show the high potential of the DeFakeNet model to discriminate between fake and actual images.

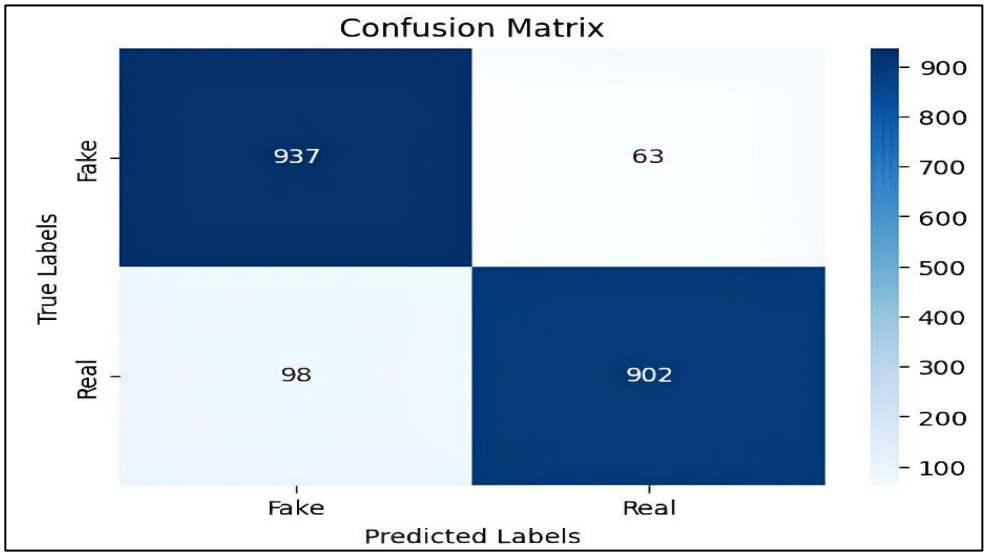

Figure 5. Confusion Matrix

The research examines the capacity of the proposed model to identify authentic and fake images through the ROC curve along with its AUC score derived from the predicted probabilities. Figure 6 (left) presents the ROC curve for the robust model. The corresponding AUC score is as high as 97.64%, which is evidence of a strong model. In fact, the curve is very close to the top-left corner, indicating that the model is very successful in terms of both the reduction of false negatives and the increase of true positives. The best ROC curve can be in the top-left area where both Specificity and sensitivity are maximum.

Precision-Recall Curve: The research measures the DeFakeNet model by means of a Precision-Recall (PR) curve. The PR score along different thresholds represents the exchange between precision and recall, as depicted in Figure 6 (right). Very good recall can be achieved together with high precision by the model if the PR score is high. A parameter value close to 1 demonstrates that the DeFakeNet model is very effective. As result, an excellent PR score of 97.40 was achieved in the task of deepfake image detection, which is very instrumental in pushing the frontier of this area further.

Matthews Correlation Coefficient (MCC): The accurate detection of a binary classification model is determined by MCC. It is a measure that balances true and false positives and negatives, making it very suitable for many image data.

DeFakeNet yielded an MCC score of 83.95%, demonstrating excellent agreement between the predicted and true classes. The model seems to be able to distinguish both classes of face images quite accurately, thus demonstrating its competence to variations in the dataset.

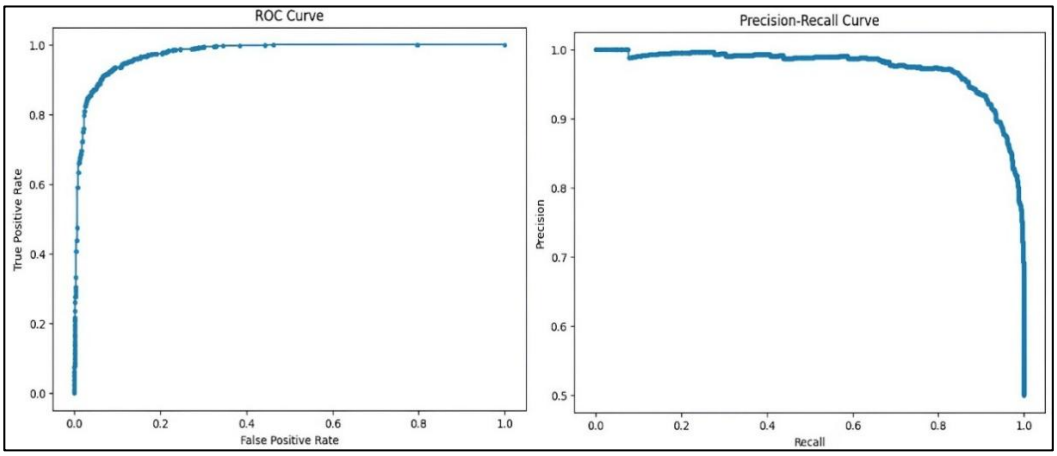

Figure 6. ROC Curve (Left) and PR Curve (Right) for DeFakeNet

4.1 Authentication of Images via Novel DeFakeNet

The images in Figure 7 have been annotated with the labels predicted by the model. Through visual inspection of the detection results, one can observe the model's effectiveness on unseen faces. In everyday scenarios, the model showcases its applicability by delivering accurate and dependable image classification.

Figure 7. Prediction on Unseen Images using Proposed DeFakeNet Model

4.2 Comparison with Related Research Works

This research compares and analyses the proposed model's performance with several research findings documented in Table 4, based on the RVF10K dataset, which shows that the DeFakeNet model emerged as the most suitable for identifying deepfakes, exceeding the performance of other models.

Table 4. Comparison Study of DeFakeNet with Recent Existing Research

Model	Train Accuracy (in %)	Test Accuracy (in %)
LeNet [21]	97.1	75
VGG16 [21]	95	50

ResNet50 [22]	85.3	82.6
DenseNet121 [23]	88.7	85.4
InceptionV3 [24]	90.5	86.9
EfficientNetB0 [25]	87.1	80.3
ViT-B/16 [26]	92.4	88.7
Swin Transformer [27]	94.2	91.2
DeFakeNet	98.88	91.95

The primary metric for comparison across all the existing studies is 'Test Accuracy', as this was the most consistently provided metric in the cited papers [22-27]. A more detailed metrics based comparison was also prepared (Table 5). However, these specific metrics were not uniformly reported in all the reviewed studies. Table 5, therefore compares the models for which the data were available [21].

Table 5. Detailed Metric Comparison with Recent Studies

Model	Precision (in %)	Recall (in %)	F1-Score (in %)
LeNet [21]	75	75	75
VGG16 [21]	25	50	33
DeFakeNet	93.47	90.20	91.81

In 2024, Pathak et al. [21] implemented many deep learning models among which LeNet and VGG16 performed poorly, failing to identify fake face images. The models implemented by [21] were pre-trained, and the lack of model tuning results in such low performance. Several baseline models involving ResNet50 [22], DenseNet121 [23], InceptionV3 [24], EfficientNetB0 [25], ViT-B/16 [26] and Swin Transformer [27] attempted to detect deepfake images. The proposed DeFakeNet model surpasses the mentioned studies in both training and testing accuracy by obtaining 98.88% training accuracy and 91.95% test accuracy on unseen data in comparison to these existing models. In addition to this, Table 5 shows that DeFakeNet maintains a strong, balanced performance across all metrics Precision (93.47%), Recall (90.20%), and F1-Score (91.81%), outperforming the other models.

5. Conclusion

The DeFakeNet model was created as a result of creative research to address the increasing problem of identifying realistic deepfake images generated by artificial intelligence. It uses a structured pipeline with data preprocessing, training, and evaluation against an equal number of real and fake images. The proposed model performs exceptionally well, attaining 91.95% accuracy and an impressive ROC AUC score of 97.64%. The limitation of the present research study is that the proposed model was trained and tested on one large dataset comprising 10,000 high-quality images. Although the ROC AUC is notably high, the performance is validated on a split of unseen tests from the dataset. Another limitation is that this model was primarily tested on StyleGAN-generated faces. Further research is needed to confirm the model's resilience across a wider range of generative methods. It could also focus on investigating more complex components that could result in improved performance, including

the use of transformers or an attention mechanism. While this model presents promising results, it will serve as a foundational model for more computationally intensive models that can effectively address emerging deepfake threats.

Acknowledgement

This research has received no external funding grant, and this is self-funded research. Debasish Samal wrote the main manuscript text along with model concept, design, figures, training, testing and validation. The critical review and analysis of the manuscript was done by Dimple Nagpal, Prateek Agrawal, Vishu Madaan and Wou Onn Choo.

Ethical Statement

Ethical approval was not needed for this research, as it utilised publicly available image set which did not contain patient information. The fake face images in the manuscript, are not of real people or taken directly from any website, so consent is not necessary.

Data Availability

The 'DeFakeNet' model architecture and associated coding files developed for this research will be accessible to interested researchers upon reasonable request to the author responsible for design and development of this research (Debasish Samal). The proposed model training and dataset samples and evaluation utilise real and fake face images sourced from dataset publicly available 'Real vs Fake Faces Balanced Dataset with Multiple Dataset Splits' which is a restructured version of existing 'RVF10K (Real vs Fake Faces 10k)' dataset from Kaggle.

The RVF10K dataset is licensed as per the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC-BY-NC-SA 4.0). The license enables us to utilise, modify, and distribute the dataset for non-commercial research purposes, provided that the original authors receive proper acknowledgement and the dataset is not used for business purposes. According to these conditions, the image dataset has been restructured into multiple training, testing and validation sets and is freely available to download online on Zenodo to facilitate reproducibility, transparent, seamless and comparable integration into future deepfake detection research.

Kunichetty, S. (2023). Real vs Fake Faces – 10k (RVF10K) [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/sachchitkunichetty/rvf10k

Samal, D., Agrawal, P., & Madaan, V. (2024). Real vs Fake Faces Balanced Dataset with Multiple Dataset Splits (Version 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14532968

Conflict of Interest

The authors state that there are no conflicts of interest in relation to this paper. The research was carried out solely for academic reasons, and no financial or personal connections affected the results or conclusions reported in this research.

References

[1] Kim, Eunji, and Sungzoon Cho. "Exposing Fake Faces Through Deep Neural Networks Combining Content and Trace Feature Extractors." IEEe Access 9 (2021): 123493-123503.

[2] Kosarkar, Usha, Gopal Sarkarkar, and Shilpa Gedam. "Revealing and Classification of Deepfakes Video's Images Using a Customize Convolution Neural Network Model." Procedia Computer Science 218 (2023): 2636-2652.

[3] Abu-Ein, Ashraf A., Obaida M. Al-Hazaimeh, Alaa M. Dawood, and Andraws I. Swidan. "Analysis of the Current State of Deepfake Techniques-Creation and Detection Methods." Indonesian Journal of Electrical Engineering and Computer Science 28, no. 3 (2022): 1659-1667.

[4] Ben Aissa, Fatma, Monia Hamdi, Mourad Zaied, and Mahmoud Mejdoub. "An Overview of GAN-DeepFakes Detection: Proposal, Improvement, and Evaluation." Multimedia Tools and Applications 83, no. 11 (2024): 32343-32365.

[5] Krueger, Natalie, Mounika Vanamala, and Rushit Dave. "Recent Advancements in the Field of Deepfake Detection." arXiv preprint arXiv:2308.05563 (2023).

[6] KV, Daya Sagar, D. B. K. Kamesh, T. Srinivasa Rao, and Chinta Venkata Murali Krishna. "Detecting Fake Faces in Smart Cities Security Surveillance Using Image Recognition and Convolutional Neural Networks." ECS Transactions 107, no. 1 (2022): 19749.

[7] Zhang, Mingxu, Hongxia Wang, Peisong He, Asad Malik, and Hanqing Liu. "Improving GAN-Generated Image Detection Generalization Using Unsupervised Domain Adaptation." In 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, 1-6.

[8] Sharma, Jatin, Sahil Sharma, Vijay Kumar, Hany S. Hussein, and Hammam Alshazly. "Deepfakes Classification of Faces Using Convolutional Neural Networks." Traitement du Signal 39, no. 3 (2022).

[9] Wen, Lilong, and Dan Xu. "Face Image Manipulation Detection." In IOP conference series: materials science and engineering, vol. 533, no. 1, IOP Publishing, 2019, 012054.

[10] Dhar, Arpita, Prima Acharjee, Likhan Biswas, Shemonti Ahmed, and Abida Sultana. "Detecting Deepfake Images Using Deep Convolutional Neural Network." PhD diss., Brac University, 2021.

[11] Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection." Information Fusion 64 (2020): 131-148.

[12] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." Communications of the ACM 63, no. 11 (2020): 139-144.

[13] Wang, Zhikan, Zhongyao Cheng, Jiajie Xiong, Xun Xu, Tianrui Li, Bharadwaj Veeravalli, and Xulei Yang. "A Timely Survey on Vision Transformer for Deepfake Detection." arXiv preprint arXiv:2405.08463 (2024).

[14] Petmezas, Georgios, Vazgken Vanian, Konstantinos Konstantoudakis, Elena EI Almaloglou, and Dimitris Zarpalas. "Video Deepfake Detection Using A Hybrid CNN-LSTM-Transformer Model for Identity Verification." Multimedia Tools and Applications (2025): 1-20.

[15] Suratkar, Shraddha, and Faruk Kazi. "Deep Fake Video Detection Using Transfer Learning Approach." Arabian Journal for Science and Engineering 48, no. 8 (2023): 9727-9737.

[16] Tunguz, Bojan. “140k Real and Fake Faces.” Kaggle, 2019. https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

[17] Tunguz, Bojan. “1 Million Fake Faces.” Kaggle, 2020. https://www.kaggle.com/datasets/tunguz/1-million-fake-faces.

[18] Arnaud58. “Flickr-Faces-HQ Dataset (FFHQ).” Kaggle, 2019. https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq.

[19] Kunichetty, Sachchit. “Real vs Fake Faces – 10k (RVF10K).” Kaggle, 2023. https://www.kaggle.com/datasets/sachchitkunichetty/rvf10k.

[20] Samal, Debasish, Pooja Agrawal, and Vikas Madaan. “Real vs Fake Faces Balanced Dataset with Multiple Dataset Splits (Version 1.0).” Zenodo, 2024. https://doi.org/10.5281/zenodo.14532968

[21] Pathak, Lakshin, Mili Virani, Mohammad S. Obaidat, Prachita Patel, Jigna J. Hathaliya, Nilesh Kumar Jadav, Sudeep Tanwar, and Nagendar Yamsani. "A Transfer Learning-Based Intelligent Judiciary System for Public Safety." In 2024 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), IEEE, 2024, 1-6.

[22] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.

[23] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 4700-4708.

[24] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going Deeper with Convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, 1-9.

[25] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks." In International conference on machine learning, PMLR, 2019, 6105-6114.

[26] Dosovitskiy, Alexey. "An Image is Worth 16x16 Words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[27] Kim, Chaewon. "Distinguishing AI-Generated and Real Images Using Swin-Transformer." In 2025 IEEE Integrated STEM Education Conference (ISEC), IEEE, 2025, 1-6.