Pertanika J. Sci. & Technol. 34 (3): 1703 - 1725 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Enhanced Deepfake Image Detection and Classification Using Deep Learning

Debasish Samal[1], Dimple Nagpal[2], Prateek Agrawal[2,3,4]*, Vishu Madaan[3], Chhavi Sharma[5], and Wou Onn Choo[6]

[1]Department of Computer Applications, School of Computer Applications, Lovely Professional University, 144411 Phagwara, Punjab, India
[2]Department of Computer Science and Engineering, School of Computer Science and Engineering, Lovely Professional University, 144411 Phagwara, Punjab, India
[3]Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Rungta International Skills University, 490024 Bhilai, Chhattisgarh, India
[4]Department of Computer Science and Information Technology, Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia
[5]Department of Computer Science and Engineering, School of Engineering and Technology, SGT University, 122505 Gurugram, Haryana, India
[6]Department of Computer Science and Information Technology, Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia

ABSTRACT

The development of Generative Adversarial Networks (GANs) for generating realistic deepfake content through artificial intelligence brings a complex task to authenticate deepfake images. The spread of deepfakes leads to widespread distrust among people while simultaneously hurting both personal and community-based reputations through deceptive information distribution. This paper aims to develop an optimised deep learning model based on the EfficientNetV2-B1 architecture designed specifically for binary image classification of distinguishing real or fake. The proposed method has been implemented on the extensive 140k Real and Fake faces dataset from Kaggle. As other existing detection and classification approaches rely heavily on pre-trained models and a limited dataset, our model delivers compelling performance through a customised training methodology. As a result, the model was able to achieve 99.91% training and 98.76% testing accuracy coupled with the precision, recall and F1-score at 99.28%, 98.43%, 98% respectively. Furthermore, the model's performance is compared to the current techniques to show its reliability. The model's

ARTICLE INFO

Article history:
Received: 04 October 2025
Accepted: 25 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.18

E-mail addresses:
debasishsamal01@gmail.com (Debasish Samal)
dimplenagpal009@gmail.com (Dimple Nagpal)
dr.agrawal.prateek@gmail.com (Prateek Agrawal)
dr.vishumadaan@gmail.com (Vishu Madaan)
chhavibhardwaj12@gmail.com (Chhavi Sharma)
wouonn.choo@newinti.edu.my (Wou Onn Choo)
* Corresponding author

e-ISSN: 2231-8526

predictions are also interpreted using XAI visualisations, providing explainable insights into the areas of an image that contribute to its classification as either real or fake.

Keywords: AI security, computer vision, deepfake detection, efficientNetV2-B1, image forensics, peaceful society, social safety, social security

INTRODUCTION

Deepfake technology has quickly advanced to generate realistic, high-resolution media displaying individuals making false statements and acting in ways that do not truly exist. This is quite threatening to the privacy and safety of a human being. A sampling of a movie star could be produced to make it appear as if they are endorsing a brand or making offensive comments (Kim & Cho, 2021). Their lives can be disrupted because of how damaging this could be to their reputation, career, and overall credibility. Beyond the safety threat, deepfakes can be weaponised against individuals to disclose sensitive and private information, or worse, persuade someone into malicious acts. The rapid and relentless advancement of deepfake technology compels individuals and organisations to be hyper-vigilant and implement layers of security for any potential concerns associated with deepfake-related issues (Kosarkar et al., 2023). One of the key components in effective detection and classification practices related to this issue requires advanced deep learning methods (Ein et al., 2022) to counter generative AI methods implemented by GANs (Goodfellow et al., 2014), which acknowledges why this is a critical issue to address. Accurate classifications are among the most important considerations for deepfake detection and classification (Krueger et al., 2023), whether that's accurately classifying potential malignant cells in neuroimaging data or identifying directly fraudulent financial transactions.

These advanced methods anticipate and neutralise potential risks, while at the same time improving overall security measures. Without advanced detection and classification tools, subjects would be more vulnerable to scams, hacking, and possibly misdiagnosed conditions in healthcare settings. Investing in the latest technologies and consistently upgrading detection systems allows individuals to be one step ahead of risks and ultimately make it a safer environment for everyone (Kumar et al., 2023).

Rapid releases of deepfakes are abusing people’s right to privacy and causing irreparable harm to their reputations. As shown in Figure 1, the face of German-born seven-time Mr Olympia & Hollywood actor "Arnold Alois Schwarzenegger" (Right) is overlaid on the previous actor "Clint Eastwood" (Left) from the Dirty Harry movie series, all created by using a DeepFaceLab tool. This makes it hard to decipher for most audience members which photo was the original and which image was the deepfake variant.

Deepfakes are such a trend now that they are being created in rapid succession and making headlines, even today, constantly abusing people’s privacy and causing damage to

their reputation. In 2025, just before the Mahakumbh festival (the world's largest cultural and spiritual gathering that only takes place every 144 years in India), pictures of World Wrestling Entertainment (WWE) wrestlers with fake backgrounds have been disseminated, e.g., in Figure 2.

False scenarios of the Indian actor and comedian, Brahmanandam, and popular wrestlers like John Cena, Brock Lesnar, and Roman Reigns being created and depicted in AI-generated photos and videos on social media are leading to misinformation and bewilderment among fans as well as the public. The distribution of these AI-generated deepfakes reminds us of the moral quandaries that come with very advanced technologies. These image forgery cases are a glaring example of the need to be digitally literate and cautious when dealing with online content.

Figure 1. Deepfake example: Original image (left) and fake image (right)
Note. Adapted from YouTube (https://www.youtube.com/watch?v=i42HGtt5fao)

Figure 2. Viral AI-generated photos of well-known celebrities from DFRAC: John Cena and Brock Lesnar (left), Brahmanandam and Roman Reigns (right)
Note. Adapted from DFRAC (https://www.dfrac.org)

Motivation

Deepfakes hold potential terrorist threats to the truthfulness of information and could potentially sway public opinion to influence political vectors or even defame persons by way of fake materials. Keeping both society as a whole and individuals safe from the negative consequences of deepfake technology is indeed a tough mission.

The CNN model, which has been proposed in this study, proves reliable for image-related binary classification problems. It keeps its best accuracy even in practical computer vision applications like image categorisation, no matter how small its footprint is. Consequently, it can be the first step in tackling the problem of deepfake identification from various sources. As the deepfake creation process keeps on improving, the deepfake detection system that needs to be robust will also have to undergo the adaptation process.

Main Contributions

The research paper's main contributions are as follows:

- Architectural Innovation: The development of a customised and computationally efficient deep learning architecture for deepfake image detection by integrating custom convolutional blocks into the EfficientNetV2-B1 backbone. The research is utilised to replace the standard EfficientNet classification top with a custom head specifically designed to improve the model's ability to capture subtle features of a face image.
- Demonstrated real or fake Image Classification through XAI: The research presented in this study involves Explainable AI (XAI) using Gradient-weighted Class Activation Mapping (Grad-CAM). It allows the visual identification of regions within the face image, which greatly impacts the prediction of the model.
- This research follows a novel training approach for the proposed model. The core function of this approach is the synergistic employment of the Adamax optimiser with a custom 'lr_ask' callback function to adjust the learning rate dynamically. The regularisation strategy used in the study combined Global Max Pooling with specific L1/L2 regularisers and a dropout rate of 0.4 to improve model performance. This new approach results in a demonstrable and optimal balance between high accuracy and computational efficiency capable of running on moderate GPUs.
- The comparative efficiency analysis highlights key metrics performance such as Precision, Recall, and F1-Score. It also compares the proposed model with related research works to show the model's suitability for deployment on resource-limited devices, which is a critical factor for practical applications.

The rest of the research paper's framework includes a detailed examination of existing literature in the 'Literature Review' section, which provides an overview of various deepfake detection techniques and relevant studies. The paper provides detailed information

about the proposed model's structural details with the dataset used in the section 'Materials and Methods'. Section 'Results and Discussion' presents the performance results of the proposed EfficientNetV2B1 model. The study summarises its main results in the conclusion and proposes potential avenues for additional research in future studies.

LITERATURE REVIEW

The identification of manipulated media that includes synthetic images and videos created by AI calls for the detection and classification of deepfakes to be done accurately. Conventional methods allowed the processing of manually created features from photos or video movements. The fake video shows that there are detailed inconsistencies in the movement, the textures, and the facial features (Abu-Ein, 2022; Li et al., 2020).

Image-Based CNNs, for instance, are largely used for the purposes of deepfake detection and image classification, besides being engaged to solve the irregularities in movements (Zhang et al., 2022), textures (Sharma et al., 2022), and facial landmarks (Wen & Xu, 2019). At the same time, the CNN model can detect the signs of an anomaly, for instance, the inconsistent expressions or the unusual textures that are a result of the deepfake process. It is possible to reconfigure this model for video deepfake, where it looks for temporal differences between several frames (Dhar et al. 2021).

Tolosana et al. (2020) have noted the success of pre-trained autoencoders in the task of copying a group of images and comparing the copied data to a reconstructed image made by a trained model. If the reconstructed image differs from the original, it may be a modified one. The approach, by and large, is the identification of low-level artefacts in deepfakes.

The methods created by Goodfellow et al. (2020) leveraged the technology of GANs to fabricate and, at the same time, detect deepfakes. To improve the generalisation for detecting adversarially created fakes, these methods make use of a trained discriminator, referred to as 'D' in the GAN architecture and a generator 'G' to create images for distinguishing between artificially generated and authentic ones. Besides that, adversarial training is also used to amplify the performance, which means that in the training process of the model, a photo or video with substantially altered input designed to trick the model into giving incorrect results is included. This, therefore, increases the model's capability of detecting deepfakes and makes it tougher for different types of manipulation of the model.

One of the main products of this development was the creation of the Vision Transformer (ViT) model by Indigo's Enseign Meetier Lesche Infantilien lab (2024), which offers different ViT models for the detection of state-forensic crimes. Vision Transformers can identify even very small changes in the image or failures since they can represent an image's total information. Along with transfer learning, the models rely on pre-trained models on the deepfake datasets, a feature that has proven to be very useful in the domain of deepfake images.

Liang et al. (2023) decided to employ ensemble methods to improve the accuracy of detection. Thus, they merged the predictions of several models, including CNNs, RNNs, and ViTs, to get one final prediction. Ensemble methods combine a multitude of techniques (Suratkar et al., 2023) that blend the different strengths of each and, as a result, get a more effective measure against deepfakes, particularly when a wide range of manipulations is used. The effectiveness of these techniques is not constant but rather changes depending on the type of deepfake and the context. The main success is generally when the combinations of multiple methods are used, which allows for more robust protection in the face of the targeted characteristics that keep on evolving (Fatoni et al., 2025). The literature review in Table 1 gives a detailed overview of the different approaches that have been implemented, along with the description of their methodology, advantages, and drawbacks.

Table 1
Summary of deepfake detection methods

Methodology	Approach	Strengths	Limitations
Physical Face Feature Extraction	Recognises anomalies in movement, textures, and facial features (Zhang et al., 2022).	Effective for specific visual inconsistencies in manual manipulations.	Controlled scalability; not effective against sophisticated AI-created deepfakes.
Facial Appearance Attribute-based CNNs	Operates convolutional layers to spot artefacts such as inconsistent expressions or unusual textures (Dhar et al., 2021).	Robust performance in image classification identifies spatial anomalies.	Does not verify temporal information; requires extensive datasets.
Autoencoder-based Detection	Assesses original and restructured face images to identify discrepancies (Tolosana et al., 2020).	Identifies low-level anomalies; focuses on reconstruction inaccuracies.	Performance depends heavily on model size and the quality of the dataset.
GAN Discriminator-based Detection	Uses GAN discriminators to differentiate between real and artificial content (Goodfellow et al., 2020).	Generalises well to adversarial content; resilient against wide manipulation.	Computationally intensive; requires hyperparameter tuning & a less-used detection approach.
Vision Transformers (ViTs)	Encodes global image information and draws on transfer learning on pre-trained deepfake datasets (Wang et al., 2024).	Effective at identifying minor artefacts; excels at global representation of face features.	Computationally demanding; needs adjustment for specific datasets.
Ensemble Techniques	Combines CNNs, RNNs, and ViTs for improved precision (Liang et al., 2023; Suratkar et al., 2023).	Aggregates powers of different models; resistant to varied manipulations.	Increased complexity and computational needs require thorough integration.

The application of deep learning over image data has become a foundational element of modern cyber forensic investigations, providing robust frameworks for identifying digital anomalies (Awasthi et al., 2023). As the complexity of manipulations scales from static images to continuous video streams, recent 2026 breakthroughs emphasise spatiotemporal deep learning architectures such as 3DCNNs, 3DResNets, and Variational Autoencoders (VAEs) to facilitate real-time video-based deepfake detection (Agrawal et al., 2026). Concurrently, across broader computer vision domains, there is a pronounced shift towards highly optimised, lightweight image recognition models. Techniques such as architectural pruning and attention mechanisms have been successfully deployed in lightweight models like ILN-YOLOv8 to drastically reduce computational overhead while maintaining high precision (Zhou et al., 2025). These parallel advancements in cyber forensics and lightweight architectural design strongly reinforce the necessity of parameter-efficient models capable of rapid, accurate inference in resource-constrained environments.

While existing methods like ensemble models and Vision Transformers achieve high detection accuracy, they suffer from significant limitations. Large backbones like VGG19 and ResNet50 have millions of parameters and require substantial GFLOPs, making them unsuitable for mobile or real-time deployment, resulting in high computational cost. Other limitation includes models overfitting to specific artefacts in the training set.

This research addresses these gaps by proposing a lightweight EfficientNetV2-B1 architecture. Unlike existing heavy models, the proposed approach prioritises a low parameter count (~8.2M) and computational efficiency (<1.0 GFLOPs) without sacrificing accuracy, specifically targeting resource-constrained environments.

MATERIALS AND METHODS

The aimed methodology for the study is illustrated in the flow diagram shown in Figure 3. The pipeline begins with an initial pre-processing phase, where the input facial images undergo resizing and augmentation to ensure consistency. Following this, the processed images are fed into the EfficientNet-based model architecture for feature extraction and analysis. Finally, the network outputs a binary image classification, determining whether the given input is a real or fake image.

Figure 4 illustrates the input layer, which accepts 150x150 pixel face images and undertakes image normalisation to proceed with feature extraction. The proposed methodology imposes an EfficientNetV2-B1-based architecture to detect deepfake images, taking advantage of its efficient convolutional layers and inverted residual blocks to extract detailed spatial features from input images.

The Fully Connected Layer is accountable for computing the final classification scores. This layer generates the scores that inform decision-making. The Output Layer utilises these scores to produce the ultimate prediction, categorising the input image as either authentic or a forgery, with the fake one being termed as such.

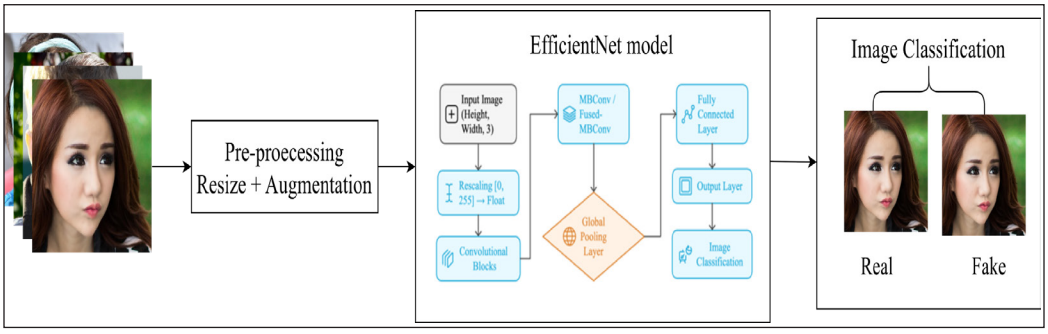

Figure 3. Proposed methodology

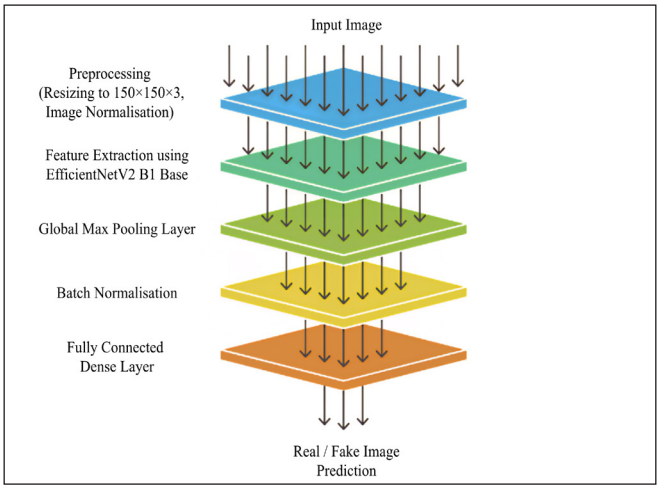

Figure 4. Stages of model implementation

Rationale for Using EfficientNet in Deepfake Detection

Both EfficientDet (Tan et al., 2020) and EfficientNet (Tan et al., 2019a) are strong neural network architectures; nevertheless, they have been designed to handle different issues. EfficientNet is a deep learning model that is mostly used in image classification tasks. It achieves this by effectively scaling model dimensions – depth, width, and resolution, using compound scaling and MBConv blocks, thereby bringing about higher accuracy and efficiency of image-level prediction tasks.

The main difference with EfficientDet is that the latter is a combination of EfficientNet as a backbone architecture with a Bi-directional Feature Pyramid Network (BiFPN) and uses this fusion for multi-scale features merging. This merges the features at different scales to allow accurate localisation and detection of objects; thus, it is a network designed for object detection & segmentation tasks. The training of EfficientNet from scratch on a big,

diverse dataset is easy, and the model converges quickly. Besides, pre-trained EfficientNet models on large datasets, such as ImageNet, provide a lever in transfer learning; thus, there is a faster convergence rate and an increased generalisation capability. The design of the model's balanced complexity is such that it avoids the problem of overfitting while it is doing so effectively on a dataset that has both real and fake images. EfficientNet is top among the choices for the deepfake detection task, and the rationale of this is that it is very accurate, it brings about efficient feature extraction, and the image classification task requirements are met through it.

EfficientNetV2-B1 Base Architecture

The EfficientNetV2 B1 model (Tan et al., 2021), one of the EfficientNetV2 family of architectures, is the basic network utilised for the work of feature extraction. The design is essentially made up of several layers of convolutions and inverted residual blocks, which allow it to capture the hierarchies of spatial and even the finest features in the input images. It is the combination of convolutional and pooling layers that has been shown in the standard architecture structure in Figure 5(a). Compared to the more complicated designs, this base model is a light, advanced deep learning architecture with 8.2 million parameters, which is aimed at lessening the complexity on relatively large datasets and is still computationally feasible for low-resource environments, such as a device with moderate GPU capabilities. Table 2 gives a summary of all the parameters and their values used in the proposed model categorically.

The design features multiple layers of convolutions, which are supplemented by such operations as Fused-MBConv (Tan et al., 2019b) for quicker execution and lower memory usage, that basically 'pull' complex spatial and hierarchical features from the input images. The fundamental measures include a first convolutional layer that is highly efficient in processing the input image. In these layers, the breakdown of standard convolution into

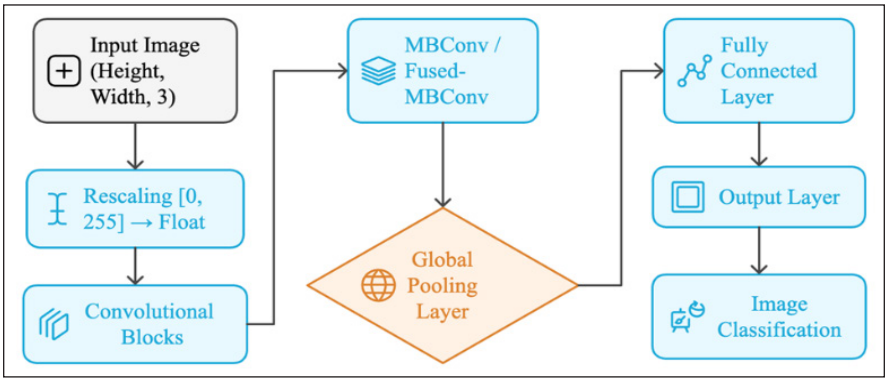

Figure 5(a). Standard structure of EfficientNetV2-B1 base model

depthwise and pointwise convolution results in substantial efficiency loss while the complexity of features is still maintained.

The suggested technique takes full advantage of EfficientNetV2-B1 features to build an accurate and computationally efficient deepfake detection model. Once the model accepts the hyperparameters, it achieves the highest accuracy of distinguishing real and fake images by employing a mixture of multi-scale features and advanced convolution blocks for capturing subtle patterns. The model is equipped with 'categorical_crossentropy' loss, optimised with 'Adamax' optimiser, further upgraded with 'lr_ask' callback function and data augmentation for better generalisation and higher accuracy.

In Figure 5(b), the standard EfficientNetV2-B1 backbone extracts features, which are then processed by the custom classification head utilising GAP, L1/L2 regularisation, and aggressive Dropout (0.4) to enhance deepfake image detection.

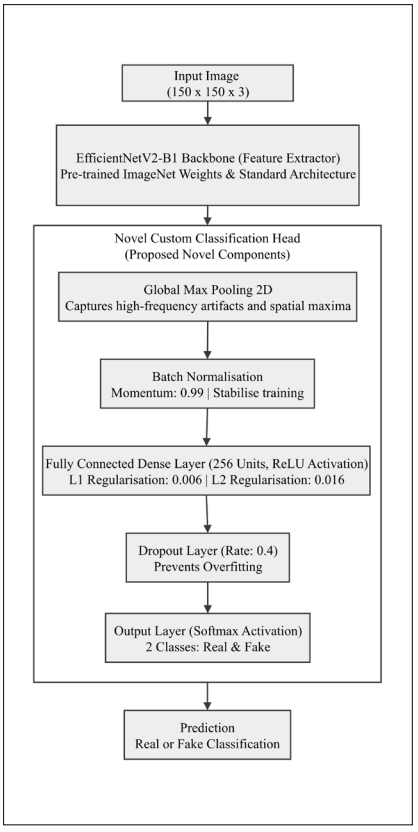

Figure 5(b). Detailed architecture of the proposed model

Table 2
Summary of hyperparameters and values

Category	Parameters	Value
Input and Architecture Specifications	Image Size	150 × 150
	Base Model	EfficientNetV2-B1
	Pooling Strategy	Global Max Pooling
	Trainable Layers	All layers (base_model.trainable=True)
	Batch Normalisation	Momentum: 0.99 Epsilon: 0.001
	Custom Dense Layer	256 Units
	Dense Layer Activation	ReLU
	Output Layer Activation	Softmax
Regularisation Techniques	Kernel Regulariser	L2 (l = 0.016)
	Activity Regulariser	L1 (l = 0.006)
	Bias Regulariser	L1 (l = 0.006)
	Dropout Rate	0.4 (Seed: 123)

Table 2 (*continued*)

Category	Parameters	Value
Training and Optimisation Parameters	Batch Size	20
	Number of Epochs	15
	Optimiser	Adamax
	Learning Rate (lr)	User-defined in a function parameter
	Loss Function	Categorical Cross-Entropy
	Metrics	Accuracy, Precision, Recall, F1- Score
	Callback: LR_ASK	Dwell: True, Factor: 0.4

Experimental Setup

The model was trained, tested, and evaluated using TensorFlow and Keras in Python on a PC equipped with an Intel Core i7 processor. The research utilised an NVIDIA GeForce RTX 3060 graphics processing unit equipped with 16 gigabytes of random-access memory.

Dataset Information

The 140k Real and Fake Faces Kaggle dataset by Tunguz (2019), which consists of 140,000 images, has been utilised in the study. Of these, 70,000 real expressions were created by NVIDIA's Flickr-Faces-HQ (FFHQ) by Rougetet (2019), and their corresponding fake expressions were created by StyleGAN from 1 million FAKE faces by Tunguz (2020). Various machine and deep learning models can be effectively trained on the look of real human faces using photos from the dataset, as depicted in Figure 6.

This dataset was chosen for the study because it includes a variety of subsets and three well-balanced core components: test, validation, and training sets. This makes it easier to construct and manage deep learning models in an organised manner, which can then be improved with new raw data. Researchers can use the dataset's CSV files, which include related metadata and annotations, for further study.

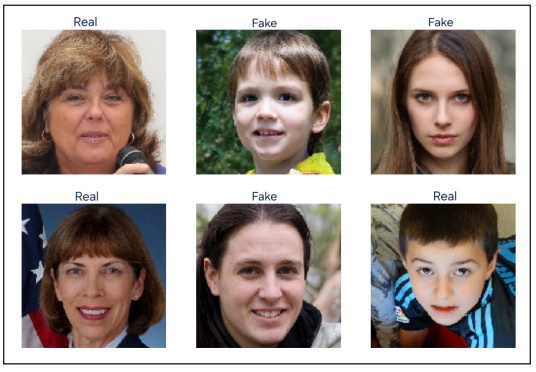

Figure 6. Real and fake image samples from the chosen dataset

In Figure 7, the bar chart illustrates the allocation of images by label in the training dataset, utilising 10,000 images for both 'fake' and 'real' categories, each in equal quantities.

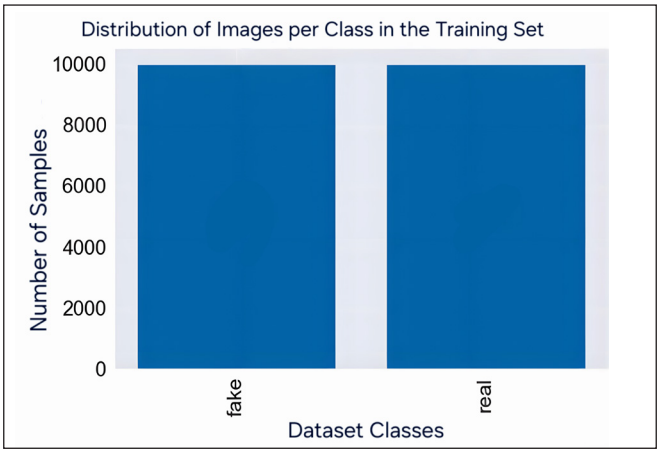

Figure 7. Balanced distribution of images in Training Set

Evaluation Metrics

The study measures model performance across various evaluation metrics, namely accuracy, recall, precision, F1 score, true positive rate (TPR), and false positive rate (FPR), after model selection and tuning.

Accuracy is defined as the number of correctly predicted instances divided by the total number of instances, as mathematically expressed in Equation 1 (Ahmed et al., 2020; Baldi et al., 2000).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad [1]$$

where TP means the number of true positives, TN is the number of true negatives, FP represents the number of false positives, and FN shows the number of false negatives.

Precision is described as the proportion of correct positive predictions to the total number of actual positive and negative cases, as mathematically expressed in Equation 2 (Iheanacho et al., 2021).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad [2]$$

Recall is defined as the proportion of actual positives that are correctly identified, calculated by dividing true positives by the total number of positives Equation 3. High recall is particularly crucial in deepfake detection, as failing to identify a synthetic image can lead to the unchecked spread of misinformation (Martin & Stevens, 2023).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [3]$$

F1 Score: The F1 score is referred to as the balanced F score due to its properties that show it as the harmonic mean of precision and recall, hence suitable for an imbalanced class Equation 4, Davis & Goadrich, 2006).

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad [4]$$

RESULTS AND DISCUSSION

Model Loss, Accuracy and F1-score at Various Training Stages

Throughout all epochs, the model showed outstanding training accuracy with minimal loss, as shown in Figure 8(a), 8(b), and 8(c). The model achieved a benchmark validation accuracy persistently exceeding 98.12%. The maximum validation accuracy reached was 98.76% at Epoch 11. Slight variations in validation loss and accuracy seen in later epoch stages indicate that the model was nearing its maximum optimal performance threshold.

Training and Validation Results

Table 3, summarises the training and validation results from Epoch 1 to Epoch 15. These results are based on the training process of the model for the deepfake detection task. For each epoch, metrics such as Training Loss, Training Accuracy, Validation Loss, Validation Accuracy, Validation Loss Improvement, Learning Rate, Next Learning Rate, and Duration in Seconds are given. By the basics of training, the model exhibited very good results in accuracy and loss decrement. Training for the first epoch, the model had a very high training loss of 3.0211 and achieved a training accuracy of 83.65%. Besides, the validation loss was 1.0067, which yielded a validation accuracy of 93.84%. A combined L1 (Lasso) and L2 (Ridge) regularisation strategy (0.006/0.016) has been employed specifically to prevent the model from overfitting to high-frequency noise artefacts common in GAN-generated images.

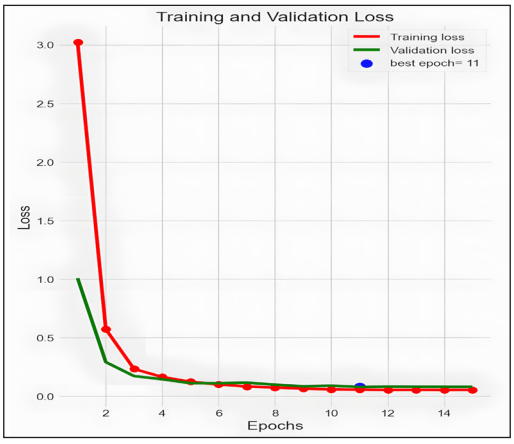

Figure 8(a). Proposed model training and validation Loss over Epochs

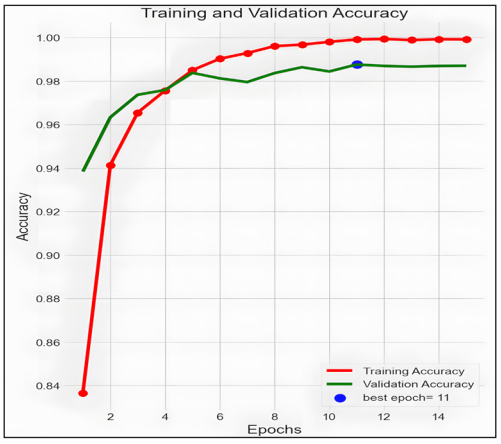

Figure 8(b). Proposed model training and validation Accuracy over Epochs

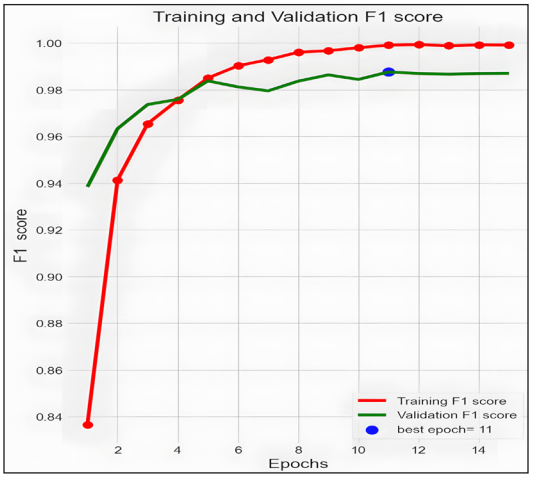

Figure 8(c). Proposed model training and validation F1-score over Epochs

The learning rate was kept at 0.001 for the first six epochs, which led to a gradual improvement of the model's performance. This trend was mainly due to the model's gradual improvement in performance until the 6th epoch, whereby it gained a training accuracy of 99.02% and a validation accuracy of 98.12%. With the start of Epoch 7, the changes in the validation loss became more pronounced and thus the loss curve started fluctuating. This sign indicated that the model was close to the point of convergence. To make the training process more stable, the training rate was lowered to 0.0004. The training accuracy rates of the model were extremely high from Epoch 11 onwards, no less than 99.88%, while only small variations were noticed in validation loss and accuracy values. The validation loss seemed to be stabilising at around 98.66% to 98.76%,

suggesting that the model was not significantly overfitting within the 15 training epochs and was thus effectively generalising.

By employing pre-trained weights from ImageNet and fine-tuning all its layers, the model manages to reveal the subtle features of fake images that are typically missed by less intricate neural networks.

Table 3
Summary of EfficientNetV2-B1 model training and validation results

Epoch No.	Train Loss	Train Accuracy	Valid Loss	Valid Accuracy	V_Loss % Improvement
1	3.0211	83.65	1.0067	93.84	0.0
2	0.5706	94.13	0.2898	96.33	71.21
3	0.2314	96.54	0.1705	97.36	41.17
4	0.1625	97.56	0.1431	97.58	16.08
5	0.1223	98.5	0.1095	98.37	23.5
6	0.0981	99.02	0.1093	98.12	0.14
7	0.0816	99.29	0.1142	97.95	-4.44
8	0.0723	99.61	0.0969	98.37	11.31
9	0.0635	99.67	0.0831	98.63	14.25
10	0.0565	99.8	0.0883	98.44	-6.23
11	**0.0546**	**99.91**	**0.078**	**98.76**	**6.13**
12	0.0526	99.93	0.0804	98.69	-3.07
13	0.0533	99.88	0.0798	98.66	-2.3
14	0.0523	99.92	0.0793	98.69	-1.57
15	0.0529	99.91	0.0792	98.7	-1.45

Note. The text in bold signifies the best results

Confusion Matrix Analysis

The efficiency of the model is thoroughly assessed using a detailed confusion matrix analysis, as shown in Figure 9, which offers insights into its precision and recall for both classes.

The machine successfully identified 9843 images as true and 9929 as false, so it is safe to say that it performed well. 71 counterfeit images were identified as true ones, while 157 authentic images were mislabeled as false. The model represents a high level of accuracy, with the largest number of true positives and true negatives supporting this. Besides, the study found 228 errors in 20000 tests, leading to an accuracy of 98.76 and an F1 score of 98.86 as depicted in Figure 10.

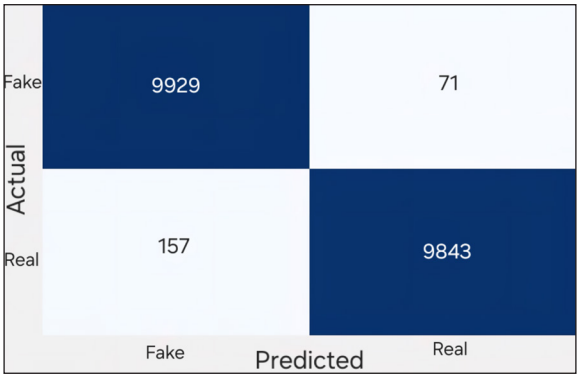

Figure 9. Confusion matrix of the EfficientNetV2-B1 model

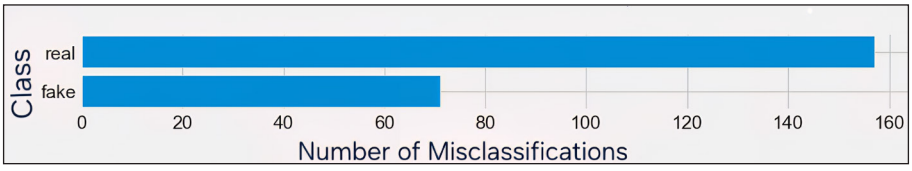

Figure 10. Classification errors occurred on Test Set by Class

Studies of the misclassification of the test dataset revealed that misclassification of real and fake samples is the main source of differences in the error rate of classification. The model tended to mistake genuine images for counterfeit the most, and about 160 were mislabelled, while only around 70 were imposter cases. The low false positives and false negatives rates pinpoint that the model is effective in reducing misclassification errors.

The confusion matrix values were used for determining the evaluation metrics of the classification model - precision, recall, and F1-score. The model was characterised by a precision of 99.28%, which means that the greatest share of the images labelled as real actually belong to the real class. The model obtained a recall of 98.43%, thus it was able to find true-positive examples among all real-world instances. The F1-score, computed as the harmonic mean of precision and recall, was 98.86%, which indicated a fair compromise between precision and recall. The combined use of these metrics gives us the conclusion that the model is a very effective instrument for separating real from artificial, with a minimal rate of misclassification.

The exceptional accuracy (98.76%) and F1-score (98.86%) achieved by the proposed model can be directly attributed to the architectural synergy between the EfficientNetV2-B1 backbone and the customised classification head. The high precision and recall are driven by the backbone's compound scaling, which efficiently captures both low-level textural anomalies and higher-level structural inconsistencies typical of GAN-generated faces. Furthermore, the robust performance is heavily influenced by the aggressive regularisation

strategy implemented. By applying Global Max Pooling in conjunction with combined L1/L2 regularisers and a steep Dropout rate (0.4), the network was explicitly penalised for memorising dataset-specific biases. This forced the model to learn generalised, high-frequency forgery artefacts rather than overfitting to the balanced properties of the 140k dataset, resulting in highly balanced predictive capabilities across both classes.

Comparative Study with Existing Models

Unlike previous research, which often relies on limited feature analysis and model tuning, the research offers a thorough analysis of how certain characteristics make predictions due to the efficient training with customised hyperparameter tuning. This study compares the proposed model's specifications along with the obtained test accuracy to previous research, demonstrating that the custom CNN model achieves higher performance in detecting deepfakes than other models.

The paper compares its approach to three existing deep learning-based methods, as described in Table 4. Sobowale et al. (2024) used VGG19 and ResNet50 models and achieved accuracy rates of 91.59% and 96.61%, respectively. Vaishnavi et al. (2024) achieved a 94% accuracy when utilising DenseNet121. Sharma et al. (2022) achieved test accuracy rates of 95.85%, 93.98%, and 86.63% using their custom CNN, ResNet50, and VGG16 models.

Table 4
Comparison of proposed model with previous deepfake detection models

Author	Model	Dataset	Accuracy (in %)	Parameter Count (M)	FLOPS (G)
(Sobowale et al., 2024)	VGG 19	140K	91.59	25.6	4.1
(Sobowale et al., 2024)	ResNet50	140K	96.61	25.58	4.09
(Vaishnavi et al., 2024)	DenseNet121	140K	94	9.4	3
(Sharma et al., 2022)	VGG 16	140K	86.63	138	15.3
(Sharma et al., 2022)	ResNet50	140k	93.98	23	4.09
(Sharma et al., 2022)	CNN	140k	95.85	11.99	2
(Tan et al., 2019a)	EfficientNetB0	RVF10K	80.30	~5.3	~0.39
(Dosovitskiy, 2020)	ViT-B/16	RVF10K	88.70	~86	~17.5
(Kim, 2025)	Swin Transformer	RVF10K	91.20	~28	~4.5
(Samal et al., 2025)	ResNet50V2	RVF10K	91.95	25.66	~4.1
(Kumar, 2024)	PViT	140K	91.92	~86	~17.5
Proposed Model	**EfficientNetV2-B1**	**140K**	**98.76**	**~8.2**	**<1.0**

Note. The dashes indicate metrics that were not uniformly reported in the original cited studies. Parameter counts and FLOPs for standard external baselines are based on approximate standard architectural values for 224×224 and 150×150 input resolutions

The major benefit of the proposed model, which is based on EfficientNetV2-B1, lies in its enhanced performance, innovative lightweight architectural structure, and detailed training procedure when compared to current deepfake detection methods. In contrast to conventional models such as VGG, ResNet, DenseNet, and basic CNNs used in previous research, EfficientNetV2-B1 offers a scalable and parameter-efficient architecture that enables deeper and accelerated learning. Larger models such as VGG16 and others may theoretically offer advantages in capturing long-range dependencies, but they are more computationally intensive, making them unsuitable for deployment on devices with limited resources. Kumar (2024) utilised a Parallel Vision Transformer (PViT) on the same 140k dataset, achieving 91.92% accuracy. The new approach achieved an accuracy of 98.76%, demonstrating that a well-tuned, lightweight CNN can outperform heavy transformer architectures on such a binary classification task while maintaining a lower computational footprint. The proposed model surpasses the performance of all previously recorded models, and it also displays high precision, recall, and F1-score, making it a highly practical and state-of-the-art solution for identifying deepfakes.

To further contextualise the performance of the proposed EfficientNetV2-B1 approach, an extended cross-dataset architectural comparison was conducted against recent state-of-the-art models, including Vision Transformers (ViTs) and the DeFakeNet architecture proposed by Samal et al. (2025). While direct 1:1 benchmarking requires identical datasets, cross-dataset evaluation provides critical insights into architectural scalability, parameter efficiency, and data dependency. For instance, heavy architectures like ResNet50V2 (25.66M parameters) and ViT-B/16 (~86M parameters) achieved 91.95% and 88.70% test accuracies, respectively, on a highly curated 10K split of the RVF10K dataset (Kunichetty, 2023). Furthermore, an older generation EfficientNetB0 baseline yielded only 80.30% on the same data. In contrast, the proposed EfficientNetV2-B1 model limits the computational burden to approximately 8.2M parameters yet successfully scales to the extensive 140k dataset to achieve 98.76% accuracy. This comparison underscores that the proposed lightweight architecture, paired with a specialised training methodology, possesses superior feature extraction capabilities when deployed across larger, more diverse datasets.

GRAD-CAM Visualisation for Deepfake Detection

This study incorporates Explainable AI (XAI) via Gradient-weighted Class Activation Mapping (Grad-CAM) to improve the interpretability and transparency of classification decisions. This entails resizing the image to a target dimension of 150x150 pixels, scaling pixel values to the range [0, 1], and appending a batch dimension to ensure compatibility with the model's input format.

As shown in Figure 11, a grad_model is generated, accompanied by a side-by-side comparison of the original image (left) and its corresponding Grad-CAM overlay.

This XAI approach facilitates not only the visual verification of the model's key areas during prediction but also enhances the reliability of the deep learning system in real-world applications (Li et al., 2020).

Figure 11. Test image prediction: Original image (Left) - Actual image label - Fake, *Grad-CAM* visualisation (Right) - Prediction - Fake (63.17%)

Ethical Considerations and Societal Impact

As deepfake technology advances, ethical concerns relating to dataset privacy, potential misuse of the technology, and algorithmic bias have become paramount. The dataset by Tunguz (2019) primarily consists of public figures or consented human faces from the web. While commonly used for research, bias in such datasets is observed, which makes the trained model difficult to generalise to unseen manipulated images. Future work must prioritise diverse, consented datasets to ensure equitable detection performance across all racial and gender groups. Evidence indicates the proposed EfficientNet model shows high potential in securing social safety by detecting forgeries, but there is also a dual-use risk where understanding detection boundaries could help adversaries generate better deepfakes. To mitigate this, the research advocates for responsible AI practices, ensuring that such tools are deployed by verified entities (e.g., social platforms, govt. cyber security bodies) rather than being open-sourced without guardrails.

CONCLUSION

In conclusion, the primary contribution of this research is the development of a highly optimised, lightweight deep learning architecture for deepfake image classification, achieved by integrating a customised classification head onto an EfficientNetV2-B1 backbone. The major findings validate the efficacy of this approach, with the model achieving an exceptional test accuracy of 98.76% and a precision of 99.28% on the extensive 140k

benchmark dataset, demonstrating robust discriminative capabilities between real human faces and AI-generated fake images. The critical practical significance of this work lies in its architectural efficiency; utilising approximately 8.2M parameters and requiring fewer than 1.0 GFLOPs, the proposed model drastically reduces computational overhead compared to massive deep residual networks and heavy transformer architectures. This structural efficiency makes the model exceptionally well-suited for real-time deployment in resource-constrained environments, mobile edge computing, and live social media filtering pipelines. While the current methodology establishes a strong baseline on a highly curated dataset, future research trajectories will prioritise extensive cross-dataset validation on uncurated, in-the-wild video frames, such as the Celeb-DF dataset. Furthermore, the integration of lightweight spatiotemporal attention mechanisms will be explored to continuously fortify digital identity authentication against increasingly sophisticated generative manipulations.

ACKNOWLEDGEMENT

This research is self-funded, and no financial support has been given for the study.

REFERENCES

Abu-Ein, A. (2022). Analysis of the current state of deepfake techniques creation and detection methods. *Indonesian Journal of Electrical Engineering and Computer Science, 28*(3), 1659–1668. https://doi.org/10.11591/ijeecs.v28.i3.pp1659-1668

Agrawal, P., Pathak, D., Madaan, V., Verma, P. K., & Choo, W. O. (2026). Spatiotemporal deep learning for real-time video-based deepfake detection using 3DCNN, 3DResNet, TCN, and VAE. *Scientific Reports, 16*, Article 18200. https://doi.org/10.1038/s41598-026-49090-1

Ahmed, J., & Ahmed, M. (2020). Ontological-based approach of integrating big data: Issues and prospects. In A. Kumar, M. Paprzycki, & V. Gunjan (Eds.), *ICDSMLA 2019* (Lecture Notes in Electrical Engineering, Vol. 601, pp. 365-378). Springer. https://doi.org/10.1007/978-981-15-1420-3_38

Awasthi, A., Das, P., Gupta, R., Varma, R., Sharma, S., Gupta, A., & Khan, H. (2023). CNN-based deep learning approach over image data for cyber forensic investigation. In *Handbook of research on thrust technologies' effect on image processing* (pp. 174-192). IGI Global. https://doi.org/10.4018/978-1-6684-8618-4.ch011

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics, 16*(5), 412-424. https://doi.org/10.1093/bioinformatics/16.5.412

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143874

Dhar, A., Acharjee, P., Biswas, L., Ahmed, S., & Sultana, A. (2021). *Detecting deepfake images using a deep convolutional neural network* [Undergraduate thesis, BRAC University]. BRAC University Institutional Repository. https://dspace.bracu.ac.bd/xmlui/handle/10361/15933

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv*. https://doi.org/10.48550/arXiv.2010.11929

Fatoni, F., Kurniawan, T. B., Dewi, D. A., Zakaria, M. Z., & Muhayeddin, A. M. M. (2025). Fake vs. real image detection using a deep learning algorithm. *Journal of Applied Data Sciences, 6*(1), 366-376. https://doi.org/10.47738/jads.v6i1.490

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27, pp. 2672-2680).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144. https://doi.org/10.1145/3422622

Iheanacho, C. O., Osoba, D. O., & Eze, U. I. (2021). Evaluation of predominant risk factors for type 2 diabetes mellitus among outpatients in two Nigerian secondary health facilities. *African Health Sciences, 21*(2), 693-701. https://doi.org/10.4314/ahs.v21i2.33

Kim, C. (2025). Distinguishing AI-generated and real images using Swin Transformer. In *2025 IEEE Integrated STEM Education Conference (ISEC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ISEC64801.2025.11147328

Kim, E., & Cho, S. (2021). Exposing fake faces through deep neural networks combining content. *IEEE Access, 9*, 123493-123503. https://doi.org/10.1109/ACCESS.2021.3110859

Kosarkar, U., Sarkarkar, G., & Gedam, S. (2023). Revealing and classification of deepfake video images using a customised convolutional neural network model. *Procedia Computer Science, 218*, 2636-2652. https://doi.org/10.1016/j.procs.2023.01.237

Krueger, N., Vanamala, M., & Dave, R. (2023). Recent advancements in the field of deepfake detection. *arXiv*. https://doi.org/10.48550/arXiv.2308.05563

Kumar, B. C. (2024). Deepfake detection using parallel vision transformers. In the *NeurIPS Workshop on Generative AI and Creativity*. https://neurips.cc/virtual/2024/98369

Kumar, L., Yadav, R. K., & Yuvaraj, S. (2023). Detecting fake faces in smart cities' security surveillance using an image. *International Journal of Scientific Methods in Engineering and Management, 1*(4), 39-48.

Kunichetty, S. (2023). *Real vs. fake faces-10k (RVF10K)* [Data set]. Kaggle. https://www.kaggle.com/datasets/sachchitkunichetty/rvf10k

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5001-5010). IEEE. https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html

Liang, B., Wang, Z., Huang, B., Zou, Q., Wang, Q., & Liang, J. (2023). Depth map guided triplet network for deepfake face detection. *Neural Networks, 159*, 34-42. https://doi.org/10.1016/j.neunet.2022.11.027

Martin, D., & Stevens, E. (2023). Statistical validation of machine learning models for diabetes prediction. In *Machine learning in healthcare conference proceedings* (pp. 110-118).

Rougetet, A. (2019). *Flickr-Faces-HQ dataset (FFHQ)* [Data set]. Kaggle. https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq

Samal, D., Nagpal, D., Agrawal, P., Madaan, V., & Choo, W. O. (2025). DeFakeNet: A ResNet50V2-based deep learning model for deepfake detection and classification. *Journal of Innovative Image Processing, 7*(4), 1356-1373. https://doi.org/10.36548/jiip.2025.4.015

Sharma, J., Sharma, S., Kumar, V., Hussein, H. S., & Alshazly, H. (2022). Deepfakes classification of faces using convolutional neural networks. *Traitement du Signal, 39*(3), 1027-1037. https://doi.org/10.18280/ts.390330

Sobowale, A. A., Adetona, B. J., Soladoye, A. A., & Omodunbi, B. A. (2024). Deepfake face recognition through modified and improved deep transfer learning. *Uniosun Journal of Engineering and Environmental Sciences, 6*(1). https://doi.org/10.36108/ujees/4202.60.0111

Suratkar, S., & Kazi, F. (2023). Deepfake video detection using transfer learning approach. *Arabian Journal for Science and Engineering, 48*, 9727-9737. https://doi.org/10.1007/s13369-022-07321-3

Tan, M., & Le, Q. V. (2019a). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105-6114). PMLR. http://proceedings.mlr.press/v97/tan19a.html

Tan, M., & Le, Q. V. (2019b). MixConv: Mixed depthwise convolutional kernels. In *Proceedings of the British Machine Vision Conference*. https://doi.org/10.48550/arXiv.1907.09595

Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10781-10790). IEEE. https://doi.org/10.1109/CVPR42600.2020.01049

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. *arXiv*. https://doi.org/10.48550/arXiv.2104.00298

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

Tunguz, B. (2019). *140k real and fake faces* [Data set]. Kaggle. https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

Tunguz, B. (2020). *1 million fake faces* [Data set]. Kaggle. https://www.kaggle.com/datasets/tunguz/1-million-fake-faces

Vaishnavi, K. D., Bindu, L. H., Sathvika, M., Lakshmi, K. U., Harini, M., & Ashok, N. (2024). Deep learning approaches for robust deepfake detection. *World Journal of Advanced Research and Reviews, 21*(3), 1629-1636. https://doi.org/10.30574/wjarr.2024.21.3.0889

Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., & Yang, X. (2024). A timely survey on vision transformer for deepfake detection. *arXiv*. https://doi.org/10.48550/arXiv.2405.08463

Wen, L., & Xu, D. (2019). Face image manipulation detection. *IOP Conference Series: Materials Science and Engineering, 533*(1), Article 012054. https://doi.org/10.1088/1757-899X/533/1/012054

Zhang, M., Wang, H., He, P., Malik, A., & Liu, H. (2022). Improving GAN-generated image detection generalisation using unsupervised domain adaptation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICME52920.2022.9859763

Zhou, X., Kan, J., Rosely, N. F. L. M., Duan, X., Cai, J., & Zhou, Z. (2025). ILN-YOLOv8: A lightweight image recognition model for crimped wire connectors. *IEEE Access, 13*, 5193-5202. https://doi.org/10.1109/ACCESS.2025.3525564