Novel Technique for a DeepFake Detection System with Transfer Learning using DenseNet-121

Akash Rajasekar
Department of Computer Science
Birla Institute of Technology and Science
Pilani, Dubai Campus
Dubai, United Arab Emirates
f20210007@dubai.bits-pilani.ac.in

Ilfa Shaheed Valiyavallappil
Department of Computer Science
Birla Institute of Technology and Science
Pilani, Dubai Campus
Dubai, United Arab Emirates
f20210048@dubai.bits-pilani.ac.in

Ria Sanjay
Department of Computer Science
Birla Institute of Technology and Science
Pilani, Dubai Campus
Dubai, United Arab Emirates
f20210025@dubai.bits-pilani.ac.in

Abstract

Deepfake detection is a vital technique to ensure the security and privacy of individuals, with the advancements in deep learning models it has become very difficult to differentiate between real and fake images. Many researchers have proposed different approaches to tackle this problem as there has been a lot of manipulated media content circulating recently which poses a potential threat of spreading misinformation. There are different existing deep learning models used by researchers for deepfake detection like CNN, RNN, and LSTM. However, this paper introduces a different approach by building a custom DenseNet-121, which is then compared with two other models, namely InceptionResNetV2, VGG16, and evaluated based on different metrics. The dataset contains both real and fake images which are all resized to obtain uniformity. This dataset is further preprocessed and tested with popularly known models for deepfake detection to provide a comparative study with respect to the proposed approach. The custom DenseNet-121 gives the best accuracy with 95.91%, performing the best compared to the InceptionResNetV2 having an accuracy of 93.4% and VGG16 with 91.5% accuracy.

Keywords: *Deepfake, Convolution Neural Network, VGG16, and DenseNet-121*

1. INTRODUCTION

In the recent times, deepfake generation technology has evolved with the advancements in deep learning models and has emerged as a challenge to media authenticity and trustworthiness. The deepfakes that are generated have become so realistic that it is almost impossible to differentiate between real and fake images. The idea was first used for entertainment and creative purposes, but off late this technology has been misused for malicious purposes which is a rising concern to security and privacy of individuals. From influencing public opinion to ruining reputations and even depicting violence, the implications of unchecked deepfake proliferation are almost all dangerous. As the threat of deepfakes grows, the development of effective detection techniques has become vital to tackle the cons.

Detection efforts are not only important for increasing the trustworthiness and authenticity of media content but also for maintaining societal stability. However, the task of detecting deepfakes poses severe challenges due to the rapid development in the field of deepfake generation methods, which continually overtakes the development of detection mechanisms. Moreover, the availability of open source deepfake tools and online tutorials make it easily accessible for malicious actors to create convincing images. Traditional approaches to deepfake detection, which rely on manual inspection and analysis of images, are ill-equipped to handle the scale and sophistication of deepfakes. Consequently, researchers have turned to machine learning and deep learning models to develop automated detection systems capable of distinguishing between real and counterfeit images. It is also vital the platforms detect deepfakes before online release as it is hard to stop the spread of deepfakes after online release.

The advent of data mining techniques have paved the way for multiple possibilities of models for deepfake detection. Many social media platforms now have multiple community guidelines which investigate the deepfake criteria as well. These detection methods make use of statistical patterns, inconsistencies in facial expressions, eye movements, and physiological signals to identify the difference between real and fake images. However, the effectiveness of this technique is dependent upon the availability of high-quality training data and robust algorithms capable of generalizing across multiple deepfake generation methodologies.

This research focuses on three different models for presenting a deepfake detection model, first a custom DenseNet-121 trained on the real and counterfeit images present in our dataset, second a VGG16 is a CNN architecture especially designed for image detection tasks. It consists of 16 convolutional layers and achieved SOTA performance on the deepfake detection task. The third model is InceptionResNetV2, which was pre-trained on the ImageNet dataset With the optimized design and efficient hyperparameter tuning, the custom DenseNet-121 achieves strong performance in deepfake detection. After testing and evaluating these models using different metrics and the best model is selected.

2. LITERATURE REVIEW

The work accomplished by Maksutov, A.A. et al. in this paper [1] focuses on the development of an efficient and accurate algorithm for detecting if deepfake technology was used to manipulate an image or a video. The model employed for the detection of deepfakes is DenseNet169, along with face distortion markers. The negative samples were developed by adding noise to the original photos using Gaussian blur, exponential blur, and Rayleigh blur. The model was evaluated on the Celeb-DF dataset, which contains the most recent deepfake videos. Each frame was analyzed for the classification of deepfake videos for an accurate result. If one percent of the frames are modified and manipulated, then the video can be classified as a deepfake. The same can be said for images with multiple faces. The AUC score for the proposed model with Rayleigh blur is 60.1 percentile when compared to the three best-proposed methods. However, the research must surely continue due to the continuous evolution of Deepfake algorithms.

Patel, M. et al. [2] intend to intercept the increased use of a new application with the help of Generative Adversarial Networks (GANs), i.e., deepfakes, that threaten the privacy and security of an individual’s data. The interception is done through the use of transfer learning; the proposed model has three phases. The first phase includes face detection and extraction using the MTCNN (Multi-task Cascaded Convolutional Networks) model from the facenet-pytorch library, the dataset was extracted from Kaggle, i.e., the Deepfake Detection Challenge. The second phase involves feature extraction with the help of VGG16, ResNet50, MobileNet, DenseNet, and InceptionV3. The third phase is classification using Random Forest (Ensemble learning), which provides the advantage of high accuracy without overfitting. The accuracy of the proposed model is 90.2%, with MobileNet showing good results as the best different feature extractor after the application of numerous evaluation metrics.

In another study, Zhu, K., Wu, B., and Wang, B. [3] makes use of open-source algorithms to create videos capturing the unique, distinctive features of a deepfake video. In recent years, we have faced an emerging threat of swapped faces in videos using AI, whereas companies are investing heavily in generating a good dataset for the detection of deepfakes. The integration of classification using a clustering-based embedded regularization term is used to improve the efficiency of the Xception network model, which contains a minimum number of parameters. Three datasets are employed in this model, i.e., UADFV, Celeb-DF, and DeepFakeDetection. UADFV is a low-quality video dataset, whereas Celeb-DF and DeepFakeDetection are high-quality video datasets. The proposed model shows an accuracy of 98.17% and an AUC score of 98.43% for a batch size of 40. The experimental evaluation metrics and comparison with additional approaches indicate the deepfake detection model shows improved accuracy and can still undergo more improvements later in the future.

In previous works, CNN was used to detect deepfakes from frames of videos by drawing out important features. However, this method can be exposed to changes in lighting, blurring, any foreign noise and the viewing angles. To avoid the error rates caused by this, Lee, G., and Kim, M. [4] have put forward a method using computer vision and DNN, to extract features from video frames. This technique is not affected by any changes in blur, degree of lighting, contrast etc. Training time has been reduced significantly in comparison with CNN as this model works on a smaller network and is powerful against any adversarial attacks. The model has been tested on the DFDC and two FF++ datasets - Face2Face and FaceSwap. OpenCV and Python were used for extraction of the features and the proposed model exhibited an accuracy of 97% for the Face2Face, 95% for the FaceSwap datasets and 96% for DFDC dataset.

Generative adversarial networks (GANs) have made the deepfake procedure convenient and easy to cultivate. The following has resulted in substantial concerns within both the private sector and government at large about the confidentiality and safety of personal identities. The conference paper proposed by Agarwal, H., Singh, A. and R. D [5] has resulted in the proposal of a novel method known as frequency domain analysis to detect anomalies unseen to the naked eye. Henceforth, the Support Vector Machine (SVM) is used as the model classifier to detect any type of forgery in an image. The fake image dataset was obtained from the generated.photos and thispersondoesnotexist.com websites.The real image dataset was acquired from

CelebA HQ and Flicker Faces. The preprocessing stage includes converting all the images to grayscale followed by a two-dimensional power spectrum using DFT (Discrete Fourier Transform). Azithumal average is further applied to acquire a dominant one-dimensional power spectrum to capture more information. The developed SVM model shows a high accuracy of 99.76% although it was trained with few samples compared to the state-of-the-art models that were fed a huge amount of data to show an improved accuracy. The framework proves beneficial for circumstances where the information or dataset is constrained. Yet the preprocessing procedure is ineffective and could potentially be improved further.

Recently proposed methods for detection of deepfakes suffer from the inability to generalize of different types of techniques used for blending videos and generative models on unseen data during the training phase. Jain, A., Korshunov, P., and Marcel, S. [6] aims to solve this attribution problem by separating each type of deepfake attack. The main goal is to not only discriminate between real and deepfake videos but also to provide a label to each type of deepfake determined at the training phase. During the test phase, the model will be used as a binary classifier provided that the training is done by balancing out the real and deepfake data. To evaluate this model, XceptionNet and EfficientNet models were taken as base, they were trained and tested as binary classifiers. These were then modified to generate an attribution on 7 different classes of videos comprising real and deepfakes of the datasets (FaceForensics++ and Celeb-DF). To enhance further, a Siamese network combined with a triplet-loss for training them to identify the above stated seven classes. The three methods (binary, attribution and triplet-loss) were tested on DeepfakeTIMIT, DF-Mobio, Google and Jigsaw datasets. Results show that TripletLoss with its NN variant works better on the Google and DF-Mobio dataset whose deepfakes are significantly different from Celeb-DF and FaceForensics++. These results show that the method of generalizing a deep model can be enhanced further on training for attribution rather than the normal binary classifier.

John, J., and Sherif, B. V. [7] provides an overview of different techniques used in detection of deepfakes, which are fake images generated using deep learning techniques that closely resemble the target individual. Two datasets were considered: one of them, named FFHQ consisted of 65,000 png images with multiple variations in features, and the other, named Deepfake Detection Challenge Dataset was composed of more than 100,000 videos, most of which were AI generated deepfakes and sourced from paid actors. The paper discusses three different methods for deepfake detection namely, feature-based, temporal-based, and deep feature-based approaches. A comparative analysis is provided which helps in understanding the pros and cons of different deepfake detection methods. Furthermore, the research proposed a semi-supervised GAN architecture which used both labeled and unlabeled data so that it acts as a multiclass classifier.

Zhang, J. et al. [8] proposes a novel ensemble technique to detect deepfakes using multiple heterogeneous features as deepfakes have turned out to be a serious threat to the privacy of individuals. Multiple existing deep learning models for deepfake generation were used to generate counterfeit images to test the working of the proposed approach. Various heterogeneous features like Gray Gradient Features, Spectrum Matrix, Contrast, Correlation, etc were extracted from both real and fake images which were then incorporated to an ensemble vector and fed into a back-propagation neural network. Dataset used for testing of the research consisted of around 5000 real images and it was split in the 4:1 ratio. This research shows that the proposed ensemble approach had outperformed the existing models for deepfake detection.

Yadav, S., Bommareddy, S., and Vishwakarma, D. K. [9] presents research which was based on the idea to improve the existing state-of-the-art techniques for deepfake detection which mostly only focuses on spatial information disregarding important temporal information. The research proposes the use of SlowFast networks, based on 3D CNNs as they take into account both spatial and temporal information. The datasets used in this research are FaceForensics++ and Kinetics-400. The proposed technique uses a SlowFast model with base architecture as ResNet, the study also shows that SlowFast models along with techniques like weighted cross entropy outperforms image based deepfake detection techniques as it addresses the unbalanced data as well.

There is an urgent need for the development of an efficient framework for detecting deepfakes due to the high rate of cybercrime and misuse of metadata, which is achieved by Raza, A., Munir, K., and Almutairi, M. [10] . A novel approach based on a hybrid of VCG16 and CNN (Convolutional Neural Network) architectures is laid out with a dataset, namely a benchmark deepfake dataset obtained from Kaggle by the Computer Science department, Yonsei University. Transfer learning techniques such as Xception, NAS-Net, MobileNet, and VCG16 are applied for evaluation and comparison purposes. The dropout, pooling, flattening, and fully connected layers are paired to build the proposed novel model architecture utilizing the hybrid layers of VCG16 and CNN. The model also undergoes hyperparameter training for the development of an effective and efficient model. The model achieves an accuracy of 94% and precision of 95% and outperforms other state-of-the-art models used for comparison. The model is validated through a confusion matrix and time series analysis, The analysis predicts the minimum error rate in the recommended model.

With respect to the rapid advancement in Deep Learning models, there has been a rise in generation of over-realistic deepfakes. Ismail, A. et al. [11] proposes a new detection model which comprises the You Look Only Once (YOLO) face detector that can capture faces from video frames. A combination of two feature extraction techniques have been proposed. First using CNN that works on the Histogram of Oriented method which works exceptionally well to discriminate the spatial information in the videos. Second one using an enhanced XceptionNet CNN. The resulting set of extracted features are then combined to be fed into a series of GRUs which works on the temporal area of information between the deepfake and genuine videos. The proposed model is trained using the CelebDF-FaceForensics++ (c23) dataset and is achieved a 95.53% on AUROC score, 95.56% accuracy, 97.06% precision, 96.21% F-score, 96.21% sensitivity, and 94.29% specificity. The paper suggests that the proposed methodology can be used for detection in videos that include both video and audio modalities.

Deng, Z. et al. [12] makes use of a deep learning technique named EfficientNet-B3 and focuses on the analysis of face edge bands to identify deepfakes. This research broadly categorizes the existing deepfake detection techniques in three main types, namely, manual-feature based, biometric-feature based, and neural network feature based with the first one using frequency domain, second one using like blinking, postures, etc, and the third one using deep learning techniques to extra high-dimensional details. The dataset is split into train, test, and validation sets to improve the performance of classification. EEfficientNet-B3 is used for training and classification, and it performs well by obtaining an excellent AUC value of over 99.8% in testing results. This method also comprises data preprocessing which includes face extraction, convex hull algorithm, and dilation algorithm.

Presently, with the circulation of Deepfakes across the internet, several detection algorithms haven’t successfully identified the visual differences between a legitimate and a deepfake video. Although, such videos do leave behind forensic noise traces, which Wang, T. and Chow, K. P. [13] emphasizes on. The model proposed in NoiseDF that uses the Siamese architecture for training RIDNet denoiser which then can identify them from the face and background section.This paper also puts forward a new Multi-Head Relative-Interaction mechanism that evaluates the face with background interaction in various dimensions to accordingly classify whether it is a Deepfake or not. The model was trained on various datasets including FF++, DFDC, Celeb-DF and DeeperForensics-1.0. Results showed that the model worked exceptionally well on the DF-1.0 with AUC score above 70%. A limitation which this paper signifies is the inability in dealing with images where the face occupies a larger portion of the image which causes the face and background grids to be overlapped.

Zhang, R., Jiang, Z., and Sun, C. [14] presents a novel technique to perfect the deepfake detection model using a two-branch detection network that uses enhancements to the Xception architecture. The network structure in simple terms consists of two branches, a whole branch for detecting the complete video and a local branch for scanning through each frame of the video. This methodology makes use of Convolutional Block Attention Module (CBAM) to improve the feature extraction mechanisms, Gated Recurrent Unit (GRU) for detecting the presence of temporal information, and different data augmentation techniques to extract features. Celeb and FaceForensics++ were two of the open datasets used in this research. Furthermore, the proposed method performs better compared to other existing methods for deepfake detection like EfficientNet, ResNet, DSP-FWA, and Mesonet.

Garg, D., and Gill, R. [15] provides an insight into the existing advanced deep learning techniques for the creation of deepfakes to alter videos and audio in a realistic manner. The authors examine the existing research on creation and detection of deepfakes, including D-CNN which is a Convolutional Neural Network used for detecting fake content. Moreover, this research highlights an inception based network model to extract features from each frame to detect deepfakes. The dataset used in this research named DefakeAVMiT is a benchmark dataset used for similar research. Furthermore, the paper presents a technique called AVoiD-DF to tackle the issue of deepfake in multiple ways like audio and video by introducing a temporal - spatial detail within the encoder and a multi modal joint decoder to combine features from various modalities.

Elpeltagy, M. et al. [16] put forward a novel deepfake video detection system to identify clones of the original video sources. The videos undergo the extraction of audio and video frames in the first phase. Each video frame and audio goes through two feature extraction techniques: XceptionNet and InceptionResNetV2. The upgraded XceptionNet model extracts spatial features from the video frames, resulting in a feature representation for the frames. The improved InceptionResNetV2 is based on the Constatnt-Q Transform technique (CQT). It helps to gather and collect in depth time-frequency features from the audio, resulting in feature representation for audio. The resultant extracted features are combined at the mid-layer producing a bimodal information-based feature representation for the entirety of the video. The Gate Recurrent Unit (GRU) is fed with three representation level to learn and obtain significant temporal information per level. Finally, the model analyses whether the forgery is only implemented on audio, video frames, or both and provides a final result. The dataset used in this work is the FakeAVCeleb multimodal video dataset. The proposed model has a F1 score of 98.43% for detecting videos, 97.49% for detecting audio and 98.04% for detecting videos of both modalities. The proposed model has improved the efficiency of detecting deepfakes but shows a disadvantage due to the limited size of the dataset.

In efforts of adding onto the work done in [17], Wodajo, D., and Atnafu, S. [3] puts forward a Convolution Vision Transformer that aims to detect deep fakes. The architecture of the model combines a CNN and Vision Transformer network. First the model generates learnable features from the CNN model which are then used as the input to the ViT model. After learning the features, the model uses the attention method to categorize them as real or fake. The Deepfake Detection Challenge (DFDC) dataset which had around 162k images was used. DL libraries like MTCNN was used for extracting the faces. The model achieved an accuracy of 91% with an AUC score of 0.91. This model proves to work efficiently provided with just the pixel location of the images and with the nonlocal features. More importance have been given to the preprocessing of the data that is used for the training in the classification stages.

In recent years, deep fakes have proven to be a major concern to society. Creation of a deep fake image or video brings out the need for detecting its legitimacy. Despite being composed of neural networks, these images or videos do leave behind spatiotemporal traces that may be invisible to humans but easily detected by detection networks.

[18] Pashine, S. et al. compares the characteristics of four neural networks (MesoNet, ResNet-50, VGG-19, Xception) and evaluates the results based on the time for operation, rate of loss, accuracy level and how it works on different data. The datasets used are Celeb-DF and Celeb-DF-v2. Results show that for cases with no limitation on hardware resources, Xception network is the most viable. VGG-19 performs better compared to the bulkier networks but only for low to medium hardware but with a lesser accuracy. Keeping in mind the huge training and inference time taken by Xception, ResNet-50 in comparison to all the options, is evaluated as the best option. Drawing conclusions from the analyses,

the result is used in real time deployment in social media platforms.

El-Gayar, M. M. et al. in the paper [19] suggests that to keep up with the rapid advancements in technology used for the generation of deepfakes, a more complex mechanism is required for its detection. Hence, it proposes a model which works on graph neural networks (GNN). Main advantage of using graph is to correlate the content and its underneath information using graph connections. The process involves two phases - first through a mini-batch graph convolution network (GCN) and then a CNN stream which involves convolution, normalization of the batch and applying activation functions. In the final stage, flattening is done to connect the convolution and dense layers. These two network streams are fused using three different networks namely FuNet-A, FuNet-M and FuNet-C. Three different datasets were used to train the model namely: DFDC, FF++ and Celeb-DF. On the DFDC dataset, the model achieves an accuracy of 99.3% and AUC of 0.96. On FF++ it achieved an accuracy of 99.3% and on the Celeb-DF, it achieved an accuracy of 98.9% and AUC of 0.98.

The aim of Arshed, M. A. et al. [20] is to build a ViT (Vision Transformer Model) for the purpose of evaluating the feasibility of the model compared to the traditional CNN (Convolutional Neural Network) based models. The framework proposes itself as a multiple-class strategy for identifying deep counterfeit images and assesses the efficiency of accurate detection. The model encounters the obstacles of Stable Diffusion and StyleGAN2 and analyzes the prospective benefits of an accurate selection of features. The dataset is a combination of open-source data from Kaggle, an online source (thispersondoesnotexist.com), StaleDiffusion, and the StyleGAN2 encoding of Stable Diffusion. The ViT model accepts the input in the form of token sequences represented by image overlays, followed by an in-depth analysis of the images. Through dataset preprocessing, the class imbalance issue is solved by creating a similar representation for all classes. The model exhibits outstanding accuracy, precision, and recall, with an F1 score of 99.9%. The ViT model was evaluated in comparison with simple traditional CNN models (ResNet-50, VGG-16) and the former showed high performance and efficiency.

3. DATASET DESCRIPTION

This dataset is taken from Kaggle which includes 140,000 images with a combination of 70,000 real face images from the Flickr dataset and 70,000 fake images generated from Nvidia's StyleGAN. The images are resized to 256x256 pixels and split into train, validation and test datasets. The features include an id, original_path (path of the image) and label (‘real’ or ‘fake’). The size of the dataset aids in an enhanced learning to detect whether the image is a deepfake or not. [21] https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

4. METHODOLOGY

The proposed architecture aims to build three distinctive convolutional neural network (CNN) based models, namely, the InceptionResNetV2, VGG16, and custom DenseNet-121. The dataset contains three sub-folders i.e., train, valid, and test datasets, in csv format. The train and valid datasets are subject to preprocessing known as data augmentation with reference to images in the first phase. In the second phase, the model architecture is defined individually for each model, followed by model compilation and training on train and valid datasets. The third phase involves the prediction of the model on the test dataset, followed by the generation of a classification report with appropriate evaluation metrics. Finally, we compare the models based on the metrics and identify the most accurate and efficient model.

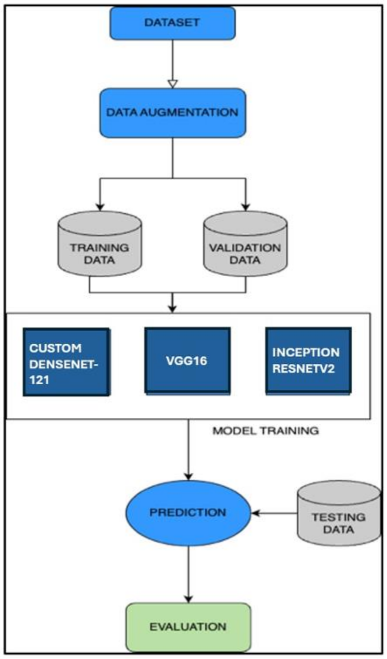


Figure 1. Architecture of the proposed model

The proposed model was run on a 12th Gen Intel(R) Core(TM) i7-12700H @2.3GB with 16GB RAMand MacOS M2 chip with 8GB RAM. The experiment is run on Python 3.11 with the model implementation done on Jupyter Notebook.

4.1 Data Analysis & Preprocessing

The dataset contains an equal number of real and fake images eliminating any imbalances in the classes. The dataset has been sampled as per the requirements with 5000, 1000 and 1000 images for train, test and valid respectively.

Data augmentation is done using TensorFlow’s Keras API. The images are augmented under various parameters:
Rescale - values of the pixels are scaled to a range [0,1].
Horizontal flip - images are flipped left to right randomly.
Rotation range - images are rotated upto 20 degrees randomly.
Zoom - images are zoomed by 20% randomly.
Brightness - image brightness is brought into the specified range.
Additionally, custom preprocessing is done by adding a random noise sample with a standard deviation of 0.1 and a mean 0 to the pixel values.

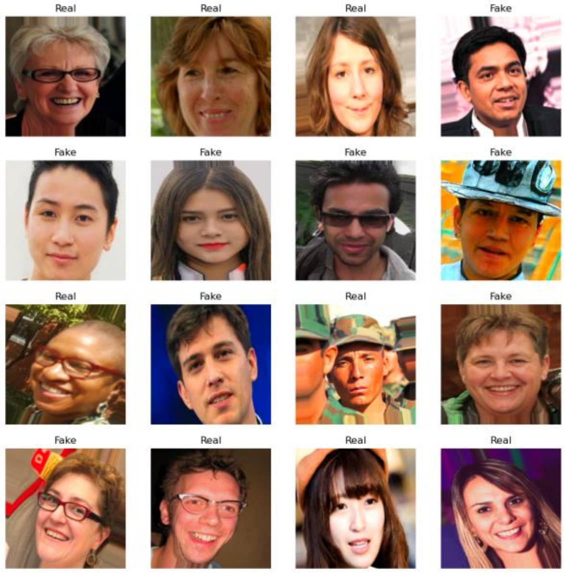


Figure 2. Dataset image analysis with predicted labels

A pixel intensity histogram depicts the pixel intensity distribution of each image in the dataset. The pixel intensity in the following histogram is represented in a distinct colour space, i.e., RGB (Red Green Blue), where the pixel values are normalized to a range of [0, 1].

It can be inferred from the following graph that the images are brighter as they are skewed towards the high intensity values, i.e., closer to the value of 1. The widespread pixel intensities indicates the presence of high contrast in images. The peaks in the histogram are tilted towards values closer to 1, representing overexposure of images. The graph also indicates images with a dynamic range that is wide and has a rich tonal range.

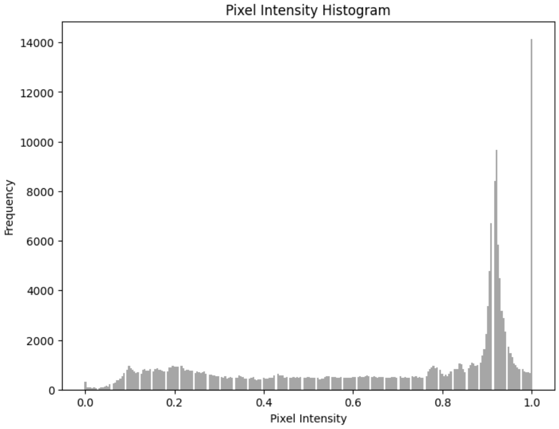


Figure 3. Graph analysing the Pixel Intensities.

4.2 Model Building

4.2.1 VGG16

The VGG16 model follows a CNN architecture which is known for its simplicity and efficiency. VGG16 consists of 16 layers that use 3x3 convolutional filters and 2x2 max-pooling layers, of the 16 layers the 13 of them are convolutional layers with the remaining 3 being fully connected layers with ReLU activations. This model is commonly used for image classification as the convolutional layers extract features from input images and identify complex patterns with increasing network depth.

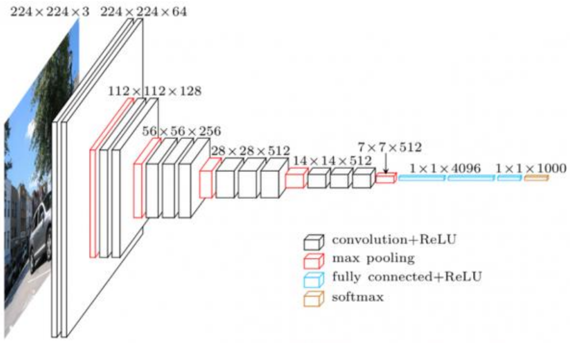
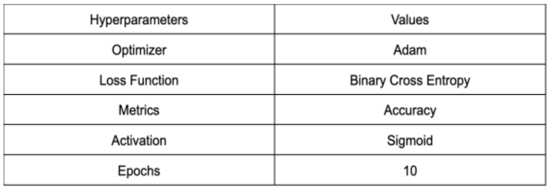


Figure 4. Architecture diagram of VGG16 model

4.2.2 Custom DenseNet-121

This model architecture is based on transfer learning with DenseNet-121 as the base model, the model is constructed with the definition of the input layer of specified shape. A global average pooling layer is included to flatten the output feature maps from DenseNet-121. After that, batch normalisation and dropout layers for regularisation are added, then a dense layer with 512 units and a ReLU activation function. For binary classification, a dense output layer with a single unit and sigmoid activation function is added. The accuracy metric, Adam optimizer, and binary cross-entropy loss function are used in the compilation of the model.



Hyperparameters	Values
Optimizer	Adam
Loss Function	Binary Cross Entropy
Metrics	Accuracy
Activation	Sigmoid
Epochs	10

Table 2. Hyperparameter analysis of the custom DenseNet-121

4.2.3 Inception ResNetV2

The InceptionResNetV2 model is initialized which was pre-trained on the ImageNet dataset. The pre-trained model's default setting is for all layers to be trainable. An additional custom classifier is applied to the base InceptionResNetV2 model. The feature maps' spatial dimensions are reduced to a vector by the GlobalAveragePooling2D() layer, after which a Dense layer with a single output neuron and a sigmoid activation function is added for binary classification.

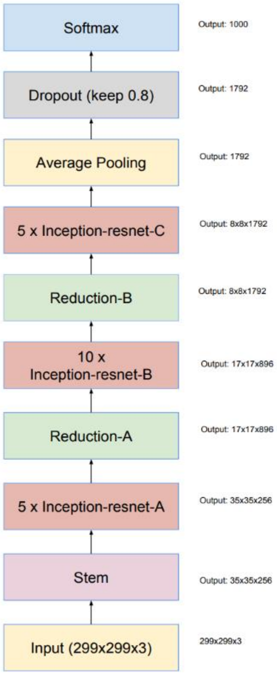


Figure 4. Architecture diagram of InceptionResNet

4.3 Model Evaluation

The performance of the model is evaluated after the training and testing phase. Since the model performs a classification task we consider the following evaluation metrics:

1. Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

2. Precision = $\frac{TP}{TP+FP}$

3. Recall = $\frac{TP}{TP+FN}$

4. F1-Score = $2 \times \frac{P \times R}{P+R}$

5. ROC curve : Plots the values of the true positive rates against the false positive rates at various thresholds.

6. AUC curve : Area under the curve is used to express the model’s ability to

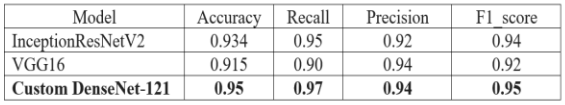
distinguish between the positive and negative samples.

7. Confusion matrix : Describes the total number of TP, FP, TN, FN predictions.

8. Specificity = $\frac{TN}{TN+FP}$

5. RESULTS AND DISCUSSIONS

The three models used were a Custom DenseNet-121, VGG16 and InceptionResNetV2. All the models were trained on the augmented dataset and evaluated using the above evaluation metrics. We find that the custom DenseNet-121 performed the best out of the three models. Below are the performance results described for each of the models.



Model	Accuracy	Recall	Precision	F1_score
InceptionResNetV2	0.934	0.95	0.92	0.94
VGG16	0.915	0.90	0.94	0.92
Custom DenseNet-121	**0.95**	**0.97**	**0.94**	**0.95**

Table 1. Performance metrics of the three models

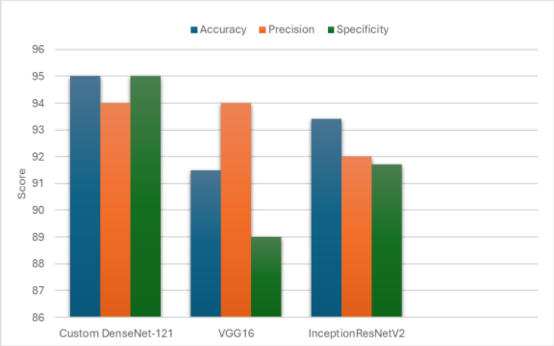
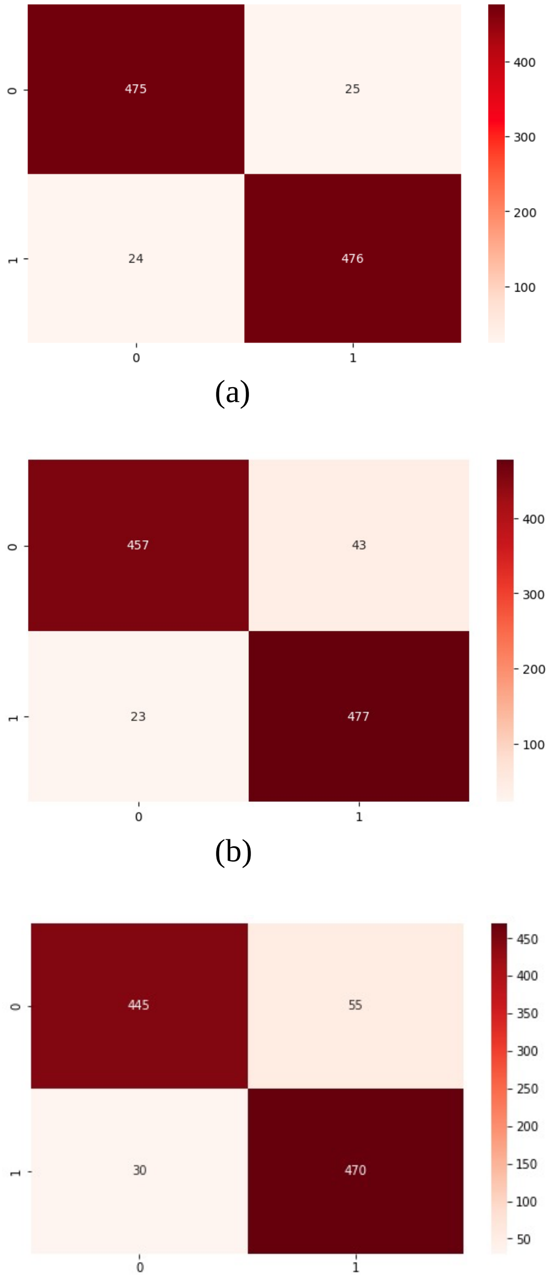


Figure 5. Comparison of the model performance

The confusion matrix of the three models are depicted below.



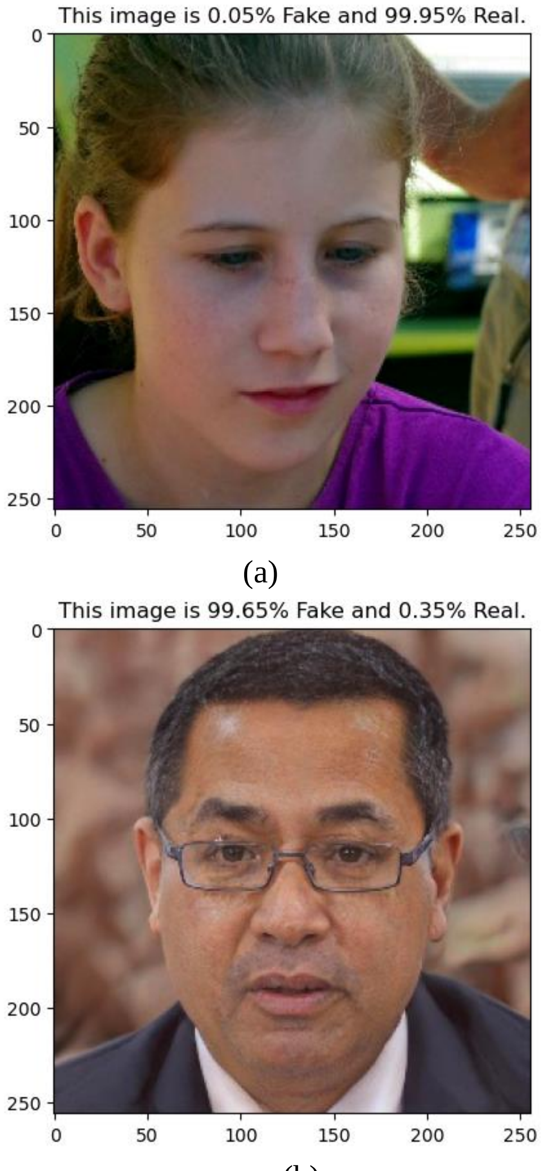
(a)

(b)

(c)

Figure 6. Analysis of the confusion matrix of the three models. (a) Confusion matrix of Custom DenseNet-121 model. (b) Confusion matrix of InceptionResNetV2 model. (c) Confusion matrix of VGG16 model.

On giving a real time test case the model describes the image as shown below

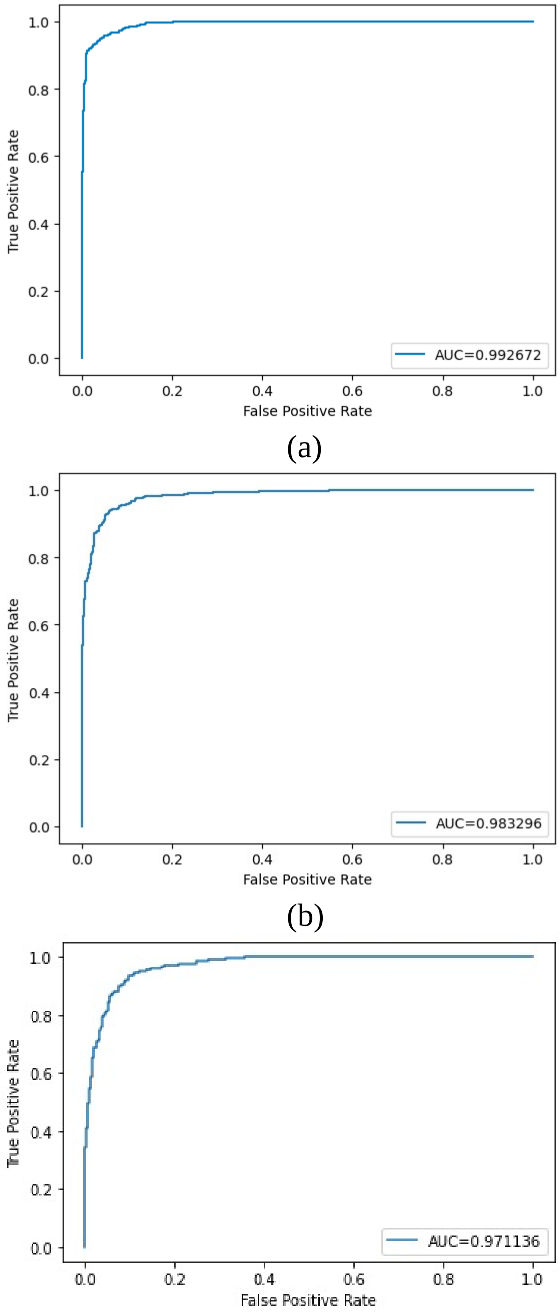

This image is 0.05% Fake and 99.95% Real.
0
50
100
150
200
250
This image is 99.65% Fake and 0.35% Real.

(a)

(b)

Figure 7. (a),(b) Real time analysis of a test image.

The ROC-AUC curves for the three models are given below.



(a)

(b)

(c)

Figure 8. The ROC-AUC curves for the three models. (a) Plot for custom DenseNet-121 model. (b) Plot for the InceptionResNetV2 model. (c) Plot for VGG16 model.

AUC (Area Under Curve) of the custom DenseNet-121 in fig (a) has a value of 0.992

which is near the value of AUC = 1.0 hence outperforms the other models. The higher AUC value depicts that the classifier can accurately distinguish between positive and negative classes.

6. CONCLUSION

A novel deepfake detection model is proposed to identify any noise in the existing images. Such an approach helps mitigate cybersecurity crimes that affect personal and commercial digital identities. The benchmark dataset 140K Real and Fake is utilized for the development of the variant CNN models. Five models, namely Custom DenseNet-121, InceptionResNetV2, and VGG16, have been trained on the dataset to classify whether the image is legitimate or a deepfake. It can be concluded that the custom DenseNet-121 model outperforms the rest of the models with an accuracy of 95%. The deployed neural networks were evaluated on various metrics and validated through a confusion matrix.

Due to the increasing deepfake inflation, there is future scope for further research to restrict the circulation of deepfakes and preserve individual digital identities. New algorithmic indicators can be investigated for accurately classifying a deepfake from a legitimate image. The model can be trained on a variety of different datasets to give a better result on unseen images and better diversity which can extend the scope of the current model. Optimization algorithms can be employed to enhance the efficiency of the deepfake detection system.

REFERENCES

[1] Maksutov, A.A. et al. (2020) "Methods of Deepfake Detection Based on Machine Learning," *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, St. Petersburg and Moscow, Russia, 2020, pp. 408-411, doi: 10.1109/EIConRus49466.2020.9039057.

[2] Patel, M. et al. (2020) “Trans-DF: A Transfer Learning-based end-to-end Deepfake Detector,” *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2020 , pp. 796-801, doi: 10.1109/ICCCA49541.2020.9250803.

[3] Zhu, K., Wu, B., and Wang, B. (2020) "Deepfake Detection with Clustering-based Embedding Regularization," *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, Hong Kong, China, 2020, pp. 257-264, doi: 10.1109/DSC50466.2020.00046.

[4] Lee, G., and Kim, M. (2021) “Deepfake detection using the rate of change between frames based on Computer Vision,” MDPI, https://doi.org/10.3390/s21217367 (accessed Mar. 3, 2024).

[5] Agarwal, H., Singh, A. and R. D, (2021) "Deepfake Detection Using SVM," *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2021, pp. 1245-1249, doi: 10.1109/ICESC51422.2021.9532627.

[6] Jain, A., Korshunov, P., and Marcel, S. (2021) “Improving generalization of deepfake detection by training for attribution,” *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2021. doi:10.1109/mmsp53017.2021.9733468

[7] John, J., and Sherif, B. V. (2022) "Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection," 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Dharan, Nepal, 2022, pp. 516-521, doi: 10.1109/I-SMAC55078.2022.9987265.

[8] Zhang, J. et al. (2022) "A Heterogeneous Feature Ensemble Learning based Deepfake Detection Method," ICC 2022 - IEEE International Conference on Communications, Seoul, Korea, Republic of, 2022, pp. 2084-2089, doi: 10.1109/ICC45855.2022.9838630

[9] Yadav, S., Bommareddy, S., and Vishwakarma, D. K. (2022) "Robust and Generalized DeepFake Detection," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-6, doi: 10.1109/ICCCNT54827.2022.9984553

[10] Raza, A., Munir, K., and Almutairi, M. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences*, *12*(19), 9820. https://doi.org/10.3390/app12199820

[11] Ismail, A. et al. (2022) “An integrated spatiotemporal-based methodology for Deepfake Detection,” *Neural Computing and Applications*, vol. 34, no. 24, pp. 21777–21791, Aug. 2022. doi:10.1007/s00521-022-07633-3

[12] Deng, Z. et al. (2022) "Deepfake Detection Method Based on Face Edge Bands," 2022 9th International Conference on Digital Home (ICDH), Guangzhou, China, 2022, pp. 251-256, doi: 10.1109/ICDH57206.2022.00046

[13] Wang, T. and Chow, K. P. (2023) “Noise based deepfake detection via multi-head relative-interaction,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14548–14556, Jun. 2023. doi:10.1609/aaai.v37i12.26701

[14] Zhang, R., Jiang, Z., and Sun, C. (2023) "Two-Branch Deepfake Detection Network Based on Improved Xception," 2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE), Changchun, China, 2023, pp. 227-231, doi: 10.1109/ICEACE60673.2023.10442716.

[15] Garg, D., and Gill, R. (2023) "Deepfake Generation and Detection - An Exploratory Study," 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gautam Buddha Nagar, India, 2023, pp. 888-893, doi: 10.1109/UPCON59197.2023.10434896.

[16] Elpeltagy, M. et al. (2023). A Novel Smart Deepfake Video Detection System. *International Journal of Advanced Computer Science and Applications*, *14*(1). https://doi.org/10.14569/ijacsa.2023.0140144

[17] Wodajo, D., and Atnafu, S. (2024) “Deepfake video detection using Convolutional Vision Transformer,” arXiv.org, http://arxiv.org/abs/2102.11126v3 (accessed Feb. 29, 2024).

[18] Pashine, S. et al. (2024) “Deep fake detection: Survey of facial manipulation detection solutions,” arXiv.org, https://arxiv.org/abs/2106.12605 (accessed Feb. 28, 2024).

[19] El-Gayar, M. M. et al. (2024) “A novel approach for detecting deep fake videos using graph neural network,” *Journal of Big Data*, vol. 11, no. 1, Feb. 2024. doi:10.1186/s40537-024-00884-y

[20] Arshed, M. A. et al. (2024). Multiclass AI-generated deepfake face detection using patch-wise deep learning model. *Computers*, *13*(1), 31. https://doi.org/10.3390/computers13010031

[21] Xhlulu. (2020). 140k real and fake faces.Retrieved from https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces