Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.029653
Article

Tech Science Press

Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods

Wahidul Hasan Abir[1], Faria Rahman Khanam[1], Kazi Nabiul Alam[1], Myriam Hadjouni[2], Hela Elmannai[3], Sami Bourouis[4], Rajesh Dey[5] and Mohammad Monirujjaman Khan[1,*]

[1]Department of Electrical and Computer Engineering, North South University, Bashundhara, Dhaka, 1229, Bangladesh
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[3]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[4]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia
[5]Department of Electronics and Communication Engineering, Brainware Group of Institutions-SDET, Barasat, Kolkata, 700124, West Bengal, India
*Corresponding Author: Mohammad Monirujjaman Khan. Email: monirujjaman.khan@northsouth.edu
Received: 08 March 2022; Accepted: 19 April 2022

Abstract: Nowadays, deepfake is wreaking havoc on society. Deepfake content is created with the help of artificial intelligence and machine learning to replace one person's likeness with another person in pictures or recorded videos. Although visual media manipulations are not new, the introduction of deepfakes has marked a breakthrough in creating fake media and information. These manipulated pictures and videos will undoubtedly have an enormous societal impact. Deepfake uses the latest technology like Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) to construct automated methods for creating fake content that is becoming increasingly difficult to detect with the human eye. Therefore, automated solutions employed by DL can be an efficient approach for detecting deepfake. Though the "black-box" nature of the DL system allows for robust predictions, they cannot be completely trustworthy. Explainability is the first step toward achieving transparency, but the existing incapacity of DL to explain its own decisions to human users limits the efficacy of these systems. Though Explainable Artificial Intelligence (XAI) can solve this problem by interpreting the predictions of these systems. This work proposes to provide a comprehensive study of deepfake detection using the DL method and analyze the result of the most effective algorithm with Local Interpretable Model-Agnostic Explanations (LIME) to assure its validity and reliability. This study identifies real and deepfake images using different Convolutional Neural Network (CNN) models to get the best accuracy. It also explains which part of the image caused the model

to make a specific classification using the LIME algorithm. To apply the CNN model, the dataset is taken from Kaggle, which includes 70 k real images from the Flickr dataset collected by Nvidia and 70 k fake faces generated by StyleGAN of 256 px in size. For experimental results, Jupyter notebook, TensorFlow, NumPy, and Pandas were used as software, InceptionResnetV2, DenseNet201, InceptionV3, and ResNet152V2 were used as CNN models. All these models' performances were good enough, such as InceptionV3 gained 99.68% accuracy, ResNet152V2 got an accuracy of 99.19%, and DenseNet201 performed with 99.81% accuracy. However, InceptionResNetV2 achieved the highest accuracy of 99.87%, which was verified later with the LIME algorithm for XAI, where the proposed method performed the best. The obtained results and dependability demonstrate its preference for detecting deepfake images effectively.

Keywords: Deepfake; deep learning; explainable artificial intelligence (XAI); convolutional neural network (CNN); local interpretable model-agnostic explanations (LIME)

1 Introduction

A Reddit user first created manipulated video clips called "deepfake" using TensorFlow [1]. The rise of deepfakes has questioned the authenticity of any digital social content. Deepfake's most widely used app is FaceApp [2], which swaps faces. For authenticity, social media platforms such as Facebook and Twitter detect and delete deep fake content. We can no longer take any video footage at its value because deepfake makes it so simple to generate false scenes. It's less about the death of truth and more about the end of faith in the truth [3]. In separate videos, the face of Argentine President Mauricio Macri has been swapped with the face of Adolf Hitler, and Donald Trump's face has taken the place of Angela Merkel's [4]. The first known attempt [5] at making a deepfake image was seen in the image of Abraham Lincoln. The Fox affiliate television station KCPQ (channel 13) broadcast a deepfake video of American president Donald Trump [6], criticizing his appearance and skin color in January 2019. Deepfake created all these fake videos containing wrong information, misleading the general public. This creates chaos in society if not appropriately addressed. Deepfakes are used to create the biggest scams, not only in videos but also in audio. In 2019 [7], the chief executive officer (CEO) of a United Kingdom (UK)-based company was tricked by audio deepfakes into transferring 220,000 euros into a Hungarian bank account. Many deepfakes include pornography of female celebrities without their permission [8]. The easy availability of deepfakes has lessened the value of human dignity. A small piece of software can generate a large number of deepfake videos that can be used to blackmail people. It becomes difficult to tell the difference between genuine and deep-fake content. So, the lives of the victims become so miserable that many of them commit suicide to compensate [9]. Fig. 1 shows deepfake images.

According to a report [10] published in 2020, more than eighty-five thousand deep fake contents were detected by December 2020. Every six months, the number of deepfake contents has doubled since the observation started in 2018.

Figure 1: Deepfake images [11]

Since the threat of deepfake has already been identified, methods for identifying deepfake are necessary. Early approaches relied on manufactured features derived from glitches and flaws in the fake video. While recent methods use Deep Learning (DL) to detect deepfake automatically. These DL models can be used to achieve tremendous accuracy, sometimes even outperforming humans. This technique has greatly benefited speech recognition, visual object recognition, object detection, and various fields [12]. DL algorithms provide unprecedented flexibility and accuracy in many security and identification domain applications. Even when trained just on statistical information, such algorithms have shown outstanding results on benchmark datasets. But despite the amazing performance of these Machine Learning (ML) algorithms, they can make mistakes too. Therefore, questions of trust and safety in ML are important [13]. Because of the growing importance of ML algorithms, the European Union enacted the General Data Protection Regulation (GDPR) [14], which includes a right to explanation. So, ML approaches should be explainable, either by utilizing inherently interpretable models or by establishing new approaches. Considering that only detection of deepfake images with better accuracy may not be enough, proper information or explanation is also necessary to understand the exact characteristics to arrive at their predictions. Thanks to the power of Explainable AI (XAI) [15], ML systems will now be able to explain their reasoning, identify their strengths and flaws, and predict their future behavior.

Deep Learning has been applied in a variety of fields, especially in recognition [16] and detection [17]. Healthcare sector is a big gun of this technique recently such in medical imaging [18], sentiment analysis [19], disease detections [20] and classifications [21] and so on [22]. Image recognition tasks have gotten tremendous hype in the Deep Learning fields. New approaches have been thinned out by more broad research platforms. Despite deepfakes being a relatively new technology, a lot of work has been published regarding deepfake classification using deep learning. Many approaches are proposed in these papers, including pairwise learning, Long Short-Term Memory (LSTM), convolutional traces, and VGG. Some researchers [23] used Generative Adversarial Networks (GAN) to detect deepfakes. The authors proposed a method that extracts a set of local features through Expectation-Maximization (EM) to detect a forensics trace hidden in images. But targeting specific features sometimes misses vital points, which could be a limitation that this work takes into consideration. Their best-performing architecture gave a validation accuracy of 90.22%. The Convolutional Neural Network (CNN) [24] is very effective and powerful in the field of image recognition and classification. This DL algorithm takes input images and

assigns weights to different attributes to tell the difference between one image and another. CNN requires less preprocessing compared to other classification algorithms. It can learn filters and characteristics after training. Many researchers preferred the CNN method due to its effectiveness in detecting deep-fake content. For example, a group of authors [25] used VGG to detect deepfakes. The image noise features are highlighted using the SRM filter layer, the face feature is augmented, and the Celeb-DF dataset is used to detect DeepFake face images.XAI model segmentation ignores image noise, which results in better classification.Their best-performing architecture gave a validation accuracy of 85.7%. DenseNet is built on a two-streamed network structure that accepts paired data as input. The researchers [26] used this pairwise learning to detect deepfakes. Not only for deepfake images, but some researchers have also worked to detect deepfake videos and audio. One of the approaches [27] proposes the idea of detecting and extracting human faces from videos and then classifying them as real or fake. They used various CNN methods, such as VGG, ResNet, and Inception, to detect deepfakes. This work compares different CNN models with respect to accuracy and loss. Their best-performing architecture gave a validation accuracy of 90.2%. Other researchers [28] implemented CNN algorithms for feature extraction from every frame in a video to train a binary classifier that learns to efficiently differentiate between real and manipulated videos. They used InceptionV3 and ResNet50 to detect deepfakes. Their best-performing architecture gave a validation accuracy of 91.56%. By increasing accuracy, the proposed model achieved an edge over the previous papers. The authors in the publication [29] applied recurrent convolutional structures to deepfake detection from audio and video. They used a vector representation of the facial region to detect deep fakes. Their best-performing architecture gave a validation accuracy of 97.83%.

DL algorithms have achieved excellent accuracy in the complex image classification and face recognition fields. But it can be detrimental to rely on a black box with apparently good performance outcomes without any quality control. Due to their nested non-linear structure, DL algorithms are usually implemented in a black-box model. These black-box models are made up straight from data [30] by an algorithm in deep learning, which means even those who develop this algorithm have no idea how variables are combined to make any predictions. There is no clear cutoff point when a model becomes a black box. Using XAI could solve this issue as complex models, such as ML or DL, with thousands or even millions of parameters, are considered “black boxes” because their behavior is difficult to explain, even when the model’s structure and weights are visible [31]. So, it is incapable of explaining decisions and actions about how they reach their predictions to human users. Hence, this lack of interpretability undermines trust in the outcomes. Explainability in DL provides a deeper understanding of how a model works and how it decides without accounting for every detail of its calculations [32]. This explainability leads to a better understanding and acceptance of the black box. XAI intends to establish more understandable models while maintaining high accuracy. XAI is becoming increasingly popular in various fields, and it is receiving increased research attention. Although few publications have been on image classification and analysis, research on strategies for interpreting more complex DL models is rapidly expanding. The authors of the paper [33] proposed a toolbox called “Neuroscope” that addresses the issue by providing state-of-the-art visualization algorithms as well as freshly modified methods for semantic segmentation of CNNs. One of the approaches [34] proposed a domain-specific explanation system for skin image analysis. The authors applied Local Interpretable Model-Agnostic Explanations (LIME) and presented the results using a Deep Neural Network (DNN) based image classifier, but for the field of deepfake, LIME is still unused. LIME [35] was one of the first two most significant efforts in the history of XAI. A tool called Lime may identify features from an image or text that are accountable for an ML model’s predictions. It is not model-specific. It can be applied to a large range of ML and DL algorithms. By feeding the comparable model inputs and watching how the predictions change, LIME tries to figure out the model’s most important features, or the major components that drive any given choice. This method provides easy explanations, such as whether the model’s predictions are driven by a specific word in a document or a feature in an image.

All of the mentioned work has achieved great accuracy, and DL algorithms have shown excellent performance. But because of the incomprehensible behavior of DL algorithms, there is a lack of liability and trust in the outcomes. Sometimes, this risk of making a wrong decision may outweigh the benefits of precision, speed, and decision-making efficacy. That is why XAI can help to understand and explain DL and neural networks better. Transparency of results and model improvement justify the adoption of XAI in the proposed method. The proposed study will use DL algorithms from CNN models to classify real and deep-fake images, compare those CNN models' accuracy, and determine which model produces the best result. This research also explains which part of the image from the dataset caused the model to make a specific classification using the LIME algorithm to assure the model's validity and reliability.

This research uses XAI to showcase the image concentration of the sample images, which is novel in detecting deepfake images with high precision. The usage of XAI makes the proposed method very reliable for detecting deepfake. Also, a comparison between different DL methods will help future researchers choose suitable methods for deepfake research, which paves the way for further improvement.

The paper is organized as follows: Method and Materials are conferred in Section 2. This section gives a gist of the system model and the whole system. Then, Section 3 demonstrates the results and analysis. Lastly, in Section 4, the paper ends with the conclusion.

2 Materials and Methodology

2.1 Dataset Description

The dataset used in this study for deepfake detection was acquired from the open-source website Kaggle [36] to train and test the model. The total number of images in the dataset is 140,000. 70,000 images contain real human faces, while the remaining 70,000 have deep faked human faces. Seventy thousand real images were retrieved from Flickr-Faces-HQ (FFHQ). FFHQ is a high-quality image dataset of human faces, created as a standard for generative adversarial networks (GAN). NVIDIA's Style-Based Generator Architecture for Generative Adversarial Networks was utilized to generate fake faces for the dataset. The images were crawled from Flickr [37]. Figs. 2 and 3 show real faces and fake faces from the dataset, respectively.

Figure 2: Real faces from the dataset

Figure 3: Fake faces from the dataset

2.2 Proposed Framework

The term "deepfake" [38] is developed from the technology "deep learning," a form of Artificial Intelligence (AI). Deep Learning (DL) and neural network technologies used to generate fake photos and videos that are difficult to distinguish from real ones are called "deepfake." Models are trained on huge

datasets to create fake content. The bigger the dataset, the more realistic the deepfakes become and the harder it is to separate them.

The proposed framework intends to detect deep fake faces and later validate the model using Explainable AI (XAI). After collecting the images, the dataset had to be split into a train, test, and validation set. The splitting was done for the training, testing, and validation sets. For better training output, the data was preprocessed. The first step in the method is to initialize to collect deepfake images, much like a good dataset. After collecting the dataset, the next step is data cleaning and data augmentation. Data augmentation is necessary because it makes the model more versatile and robust. It creates a new data scenario for the model to learn from in the training phase. By doing so, when the model faces unseen data, it can predict the output well. Data preprocessing is a big part of it also. Preprocessing the data makes things easier for the model. Precise and correct data are necessary for building a good model. After that, the proposed method has data splitting. In this phase, the main dataset is divided into train, test, and validation sets. It is important to ensure that the datasets are balanced and have an equal distribution of every data type. After splitting the dataset, the training set is used to train the model with different DL methods like InceptionResNetV2, DenseNet201, ResNet152V2, and InceptionV3. The model is validated with a validation set and a test set using different metrics like train accuracy, validation accuracy, validation loss, etc. After the training phase is finished, the model can predict an image. A sample unseen image from the test set will be taken for prediction and later verified with XAI using the Local Interpretable Model-Agnostic Explanations (LIME) algorithm.

2.3 Data Pre-Processing

To ensure the images were properly processed, they have to be checked by plotting them using Matplotlib. The images were cropped, and the face was centered. The images were rescaled to match the RGB channel.

2.4 Data Splitting

The images were resized to 256 px and later split into the train, test, and validation sets. A total of 100,000 images were taken for the train set; another 20,000 each were for validation and test sets. Both fake and real images were split so that the dataset remained balanced between the two classes. Fig. 4 shows the bar chart of the training, validation, and test sets after splitting the dataset.

2.5 Data Augmentation

The dataset was obtained through data augmentation. The augmentation process included image resizing and rescaling. Consistent results were mentioned for the batch size. The training set was shuffled for every epoch so that the model did not overfit or give repetitive results. Lastly, horizontal flips were added in the data augmentation phase to make the model more robust for unseen test images. The data augmentation was applied to the train, test, and validation sets. Lastly, fake images were labeled as 0, and real images were labeled as 1.

2.6 Pre-Trained Models

Since the dataset was ready for the training phase, Deep Learning (DL) was chosen because the models were already pre-trained with huge datasets. For the training phase, the InceptionResNetV2, DenseNet201, ResNet152V2, and InceptionV3 models from TensorFlow were picked for the analysis. The dataset was trained on these models and was closely monitored to find the best-performing model according to the training and validation accuracy score and graph.

2.7 Model Training

The system architecture provides an overview of the whole system. In this architecture, the first images from training datasets will be taken for the training phase. Using different DL methods like InceptionResNetV2, DenseNet201, ResNet152V2, and InceptionV3, the model will be trained. The system will have different layers, namely global average pooling, a dense layer with 512 neurons, an activation layer, and a classification with softmax. The average pooling layer will be flattened into a single array. Thus, the training phase will conclude. Then, taking unseen data from the test phase, the model will predict whether it is a deepfake image or a real image. Later, the same prediction will be verified using Explainable AI (XAI), which uses the Local Interpretable Model-Agnostic Explanations (LIME) algorithm. Fig. 4 gives the visual representation of the system architecture.

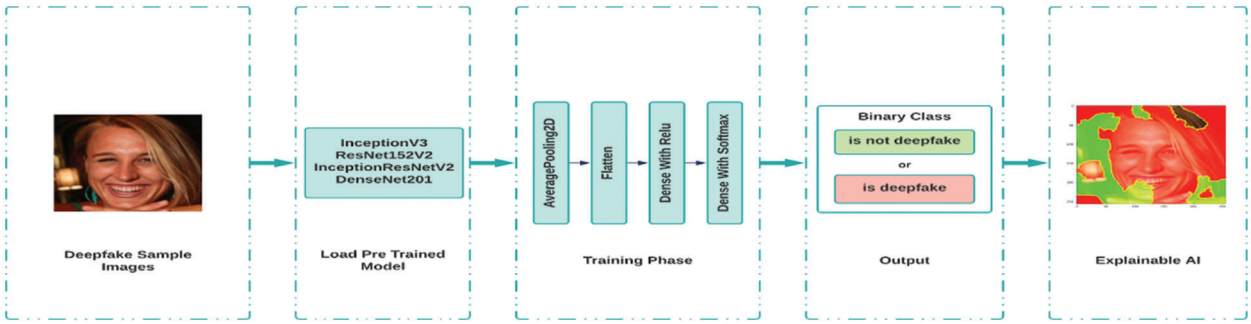


Figure 4: System architecture

The model's weights were kept the same, which was an "ImageNet" to detect the consistent comparison. The model's top layer was ignored, and the dense layer was changed accordingly for two neurons, which have fake and real classifications.

Since TensorFlow provides different Deep Learning (DL) models, every model has its own unique density and layer count. For comparison purposes, all the hyperparameters were kept the same for comparison purposes to allow proper comparison between different DL methods. None of the different models had any pre-trained top layer. The input shape of all the models was kept exactly the same as 224 px* 224 px with channel 3 for RGB format. The convolutional layer, pooling layer, Relu correction layer, and fully connected layer are different layers of the CNN model. For the proposed model, the global average pooling layer and activation layer of Relu were used with a density of 512. Additionally, for hyperparameter tuning, batch normalization was used. Model overfitting is a common issue for big datasets. To prevent this issue from happening, the proposed method implements a dropout of 0.2 before training every batch of neurons. The models were trained with the previous weight of "ImageNet."

The intention of the proposed framework is to detect deep fake faces and later validate the model using XAI. After collecting the images, the dataset had to be split into a train, test, and validation set, and the splitting was done so that the train, test, and validation set remained balanced. For better training output, the data was preprocessed. To make sure the images were properly processed, it has to be checked by plotting the images using Matplotlib.

2.8 Model Validation

For model validation, every training phase had a checkpoint where the best performing model was monitored according to validation loss and saved with a verbose of 1. This includes both the progress bars and one line for every epoch. Also, to prevent the model from learning the same findings repeatedly, the proposed model uses a reduction in the learning rate if a plateau is reached. The proposed model

monitors the plateau using validation loss with a factor of 0.2 and patience of 3. When the minimum delta reaches 0.001, the learning rate is reduced. As a result, the model only learns useful information and new patterns. The training and validation set needs to maintain steps per epoch, which are calculated by

$$Stepsperepoch = \frac{trainingsetsamplenumber}{batchsize} \tag{1}$$

In the same way, the steps for the validation set were calculated.

$$Validationsteps = \frac{validationsetsamplenumber}{validationsetbatchsize}$$

Then the transfer learning model was trained with 20 epochs.

The proposed model was validated with other different methods also. The training history was plotted using matplotlib. The history function of the training phase was stored, and the model accuracy and validation accuracy graphs were plotted and compared simultaneously. The same method was followed for training model loss and validation model loss.

In the third phase of validation of the model, the model was evaluated using the test set, which was unknown data using the model evaluation function. Finally, the validation phase concluded by verifying the model using Explainable AI (XAI).

2.9 Explainable AI

Deep learning models are considered black-box models because gaining a thorough understanding of the inner workings of deep neural networks is unyielding. Most of the time, Artificial Intelligence (AI)accepts the results without much explanation. This demands a system to explain these black-box models; thus, Explainable AI (XAI) emerged. XAI gives clarification through visualization, analysis, masking, numeric values, and the weight of features.

The Local Interpretable Model-Agnostic Explanations (LIME) algorithm was considered the best fit for the black-box model because LIME is model-agnostic, which means that it can be used for different Deep Learning (DL) models. As the proposed method went through a few comparisons between different DL methods, it was important to pick a model-agnostic XAI algorithm. The LIME algorithm uses a submodular pick to export the outcome of a model. The algorithm takes two variables, and first, using a sparse linear explanation, the algorithm explains the weights of responsible features. Secondly, the importance of these features is directly computed and later optimized using greedy algorithms. The argmax function returns the final feature weight.

Sample image reading was done through a Python module named “pillow,” where the sample image was converted, and an extra dimension was added to the array of the sample image. After that, the model had to predict the sample image using the Argmax function. For the segmentation of the sample image, the segmentation algorithm of the scikit image was used with the mark boundaries feature. The segmentation algorithm had a parameter ratio of 0.2 and a maximum distance of 200. After segmentation, the sample image was plotted using a color bar. Later, the sample image was verified using 3D masking.

To implement the LIME model and explain the black-box of deep learning, a Lime Image Explainer needs to be created. For instance, when explaining an instance, the explainer takes the sample image as an array, the predicted result of the sample image, its label amount, and the number of samples. In the proposed method, after creating the explainer, the visual analysis of the explanation had to be shown using a boundary. The image mask was verified using both positive-only and normal amounts.

3 Results and Analysis

3.1 Training Factors

Following data augmentation, the model was trained for 20 epochs with a trained generator and validation generator, resulting in good model accuracy for each model. In Tab. 1, the training factors have been summarized. The Kaggle Kernel has been used for training the models. Adam is used as an optimizer, and categorical cross-entropy is used as a loss function. Tab. 1 shows the list of materials and tools.

Table 1: List of materials and tools

Training factor	Values
Platform	Kaggle kernel
GPU	Kaggle GPU (NVIDIA Tesla K80)
Optimizer	Adam
Loss function	Categorical cross-entropy
Learning rate	0.0001
Epoch	20
Batch size	64

3.2 Model Evaluation

Every model has been trained for 20 epochs, and the accuracy and loss value of the final epoch for each model is shown in Tab. 2. After observing the table, it can be stated that every model has achieved train accuracy of 100% and their train loss value is almost zero. Besides that, every trained model has achieved decent validation accuracy. InceptionResNetV2 achieved the highest validation accuracy of 99.87% along with 0.56% of validation loss. DensNet201 also performed almost close to InceptionResNetV2. DenseNet201 achieved a validation accuracy of 99.81% along with a validation loss of 0.71%. ResNet152V2 achieved a validation accuracy of 99.19% along with a 3.93% validation loss. InceptionV3 achieved a validation accuracy of 99.68% along with a validation loss of 1.33%. InceptionV3 and ResNet152V2 have higher validation losses than other trained models. So, InceptionResNetV2 performed better than the other trained models.

Table 2: Model accuracy and loss

Model	Train accuracy	Train loss	Validation accuracy	Validation loss
InceptionV3	1.00	0.00009	0.9968	0.0133
ResNet152V2	1.00	0.00003	0.9919	0.0393
InceptionResNetV2	1.00	0.00001	0.9987	0.0056
DenseNet201	1.00	0.00001	0.9981	0.0071

To measure the performance of the trained models, all the models have been tested on unseen test data. The test set accuracy has been shown in Tab. 3. All the models have achieved decent accuracy on the test set. But compared to the other models, InceptionResNetV2 achieved the highest accuracy of 99.86%. The InceptionResNetV2 model had more trainable parameters available than the model-selected models, which had 55 million trainable parameters. Above all that, the InceptionResNetV2 model has batch

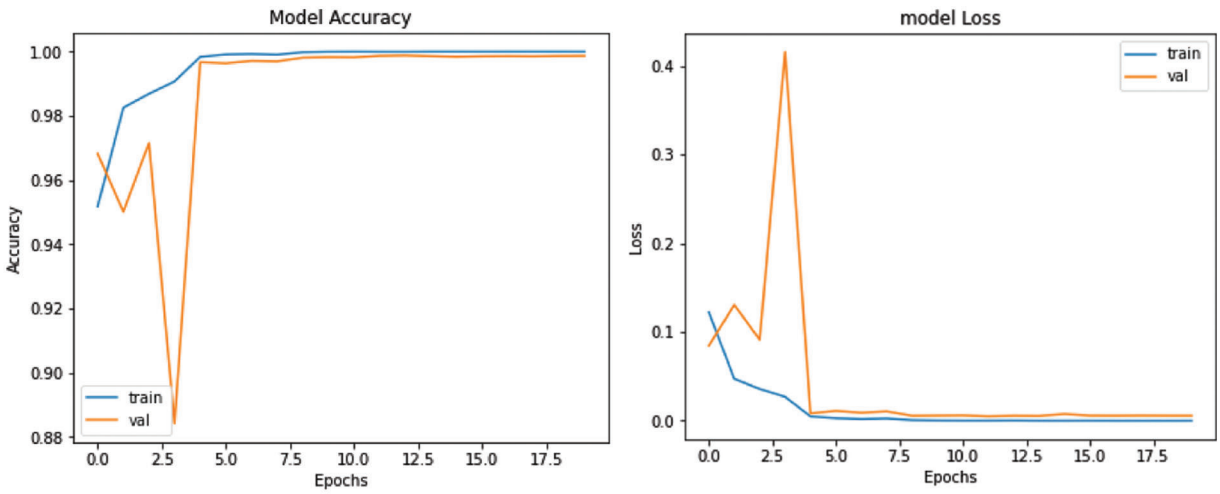
normalization, which improves the learning rate of the model. These factors lead the InceptionResNetV2 model to score higher accuracy and have a good read of the validation and test images.

Table 3: Test set accuracy

Model	Test set accuracy
InceptionV3	0.9965
ResNet152V2	0.9938
InceptionResNetV2	0.9985
DenseNet201	0.9980

In the accuracy and loss graph of every model, the accuracy rises dramatically after each epoch, and the loss decreases with each epoch. Then, after 10 epochs, every model started achieving stable accuracy and loss values for the rest of the epochs.

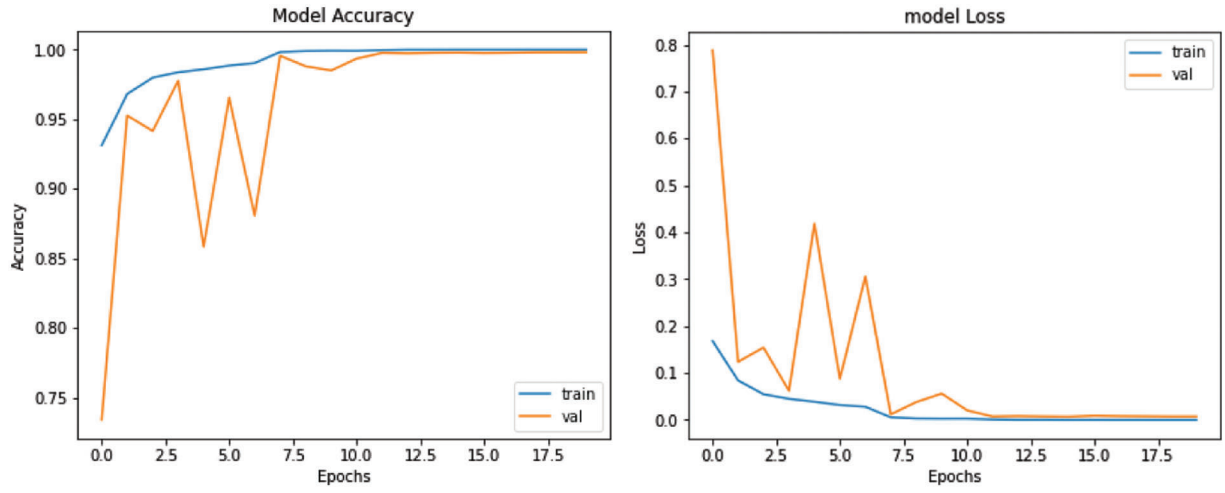
In Fig. 5, for InceptionResNetV2, the training accuracy was 95.18% in the first epoch and subsequently increased with each epoch. The model's validation accuracy was 96.82%, and it continued to rise until the final epoch. After 10 epochs, the training loss was 0.008% and the validation accuracy was 99.82%. In the final epoch, validation accuracy was 99.87% and the training loss was almost nil.



(a) (b)

Figure 5: InceptionResNetV2. (a) accuracy and, (b) loss graph

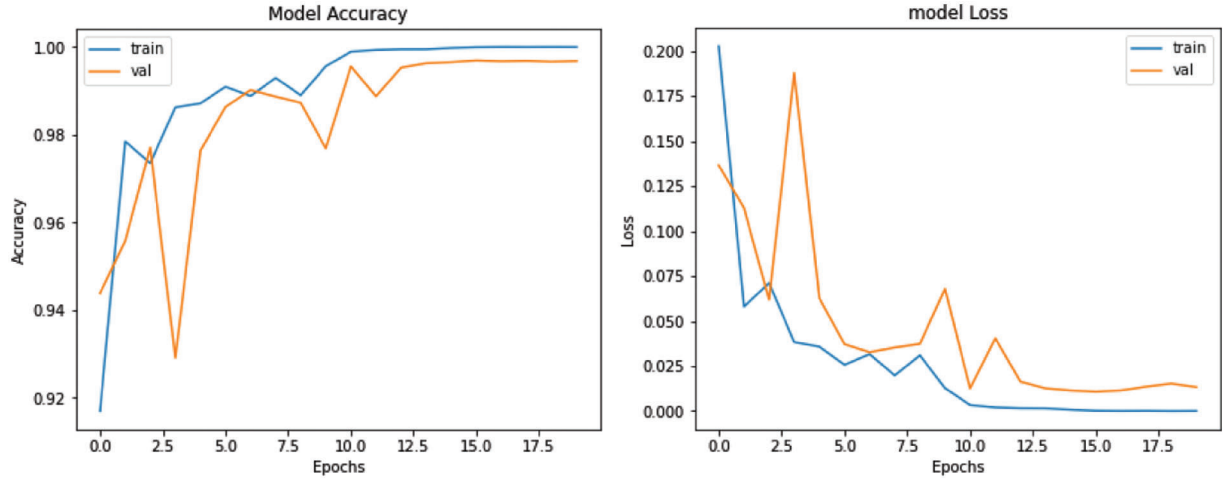
In Fig. 6, for DenseNet201, the training accuracy was 96.37% in the first epoch and subsequently increased with each epoch. The model's validation accuracy was 70.40%, and it continued to rise until the final epoch. After 10 epochs, the training loss was 0.27% and the validation accuracy was 99.35%. In the final epoch, validation accuracy was 99.81% and the training loss was almost nil.



(a) (b)

Figure 6: DenseNet201. (a) accuracy and, (b) loss graph

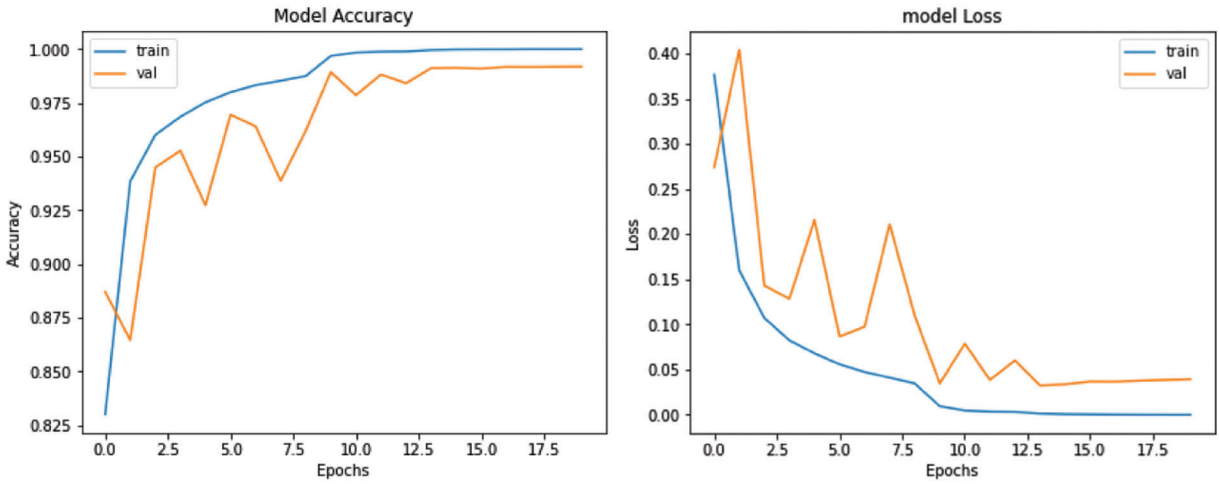
In Fig. 7, for InceptionV3, the training accuracy was 91.70% in the first epoch and subsequently increased with each epoch. The model's validation accuracy was 94.39%, and it continued to rise until the final epoch. After 10 epochs, the training loss was 1.27%, and the validation accuracy was 97.69%. In the final epoch, validation accuracy was 99.68% and the training loss was almost nil.



(a) (b)

Figure 7: IncptionV3. (a) accuracy and, (b) loss graph

In Fig. 8, for ResNet152V2, the training accuracy was 83.01% in the first epoch and subsequently increased with each epoch. The model's validation accuracy was 88.72%, and it continued to rise until the final epoch. After 10 epochs, the training loss was 0.9% and the validation accuracy was 98.94%. In the final epoch, validation accuracy was 99.19% and the training loss was almost nil.



(a) (b)

Figure 8: ResNet152V2. (a) accuracy and, (b) loss graph

3.3 Performance Validation with XAI

For deepfake analysis, it is very important to correctly identify real images to make the model trustworthy, and Explainable AI (XAI) gives Deep Learning (DL)that transparency. Artificial Intelligence (AI) needs to be ethically maintained in addressing any problems. The proposed method applied XAI to all the DL models, and of all the models, InceptionResNetV2 gave the most precise black-box explanation. For this reason, the InceptionResNetV2 model was chosen to detect deepfake and explain the model.

The Local Interpretable Model-Agnostic Explanations (LIME) algorithm was applied to the proposed model. The LIME algorithm is a popular XAI algorithm that can be used to explain image samples with visual representation. LIME is a model-agnostic algorithm that approximates the local linear behavior of the model. As a result, it can explain any model of Convolutional Neural Network (CNN) and Natural Language Processing (NLP). The visual representation of the explanation is the easiest for any human to understand. That is why the proposed model went for the visual approach.

Fig. 9 is a sample image of a human face taken to predict with the proposed model, and later it was used to explain the blackbox model with the LIME algorithm. The image has been picked from the dataset, which has not been preprocessed. The sample image is fully unseen in the proposed model. Unseen images give the best outcome in evaluating the model. Using the neural network, the model predicted that the sample image would be a deepfaked image. The model took the highest value of the neuron from the CNN method that had the highest value for the classification result. XAI was later used to unveil the reasoning for this prediction.

To use the LIME algorithm, the function must be tuned to the desired parameter. The segmentation algorithm was taken from the LIME wrapper of the scikit image, which marks boundaries along the sample image's high-weight areas. The segmentation algorithm had a ratio of 0.2 with a maximum dist of 2000. This process divides the sample image into a separate section using colorbar. Image segmentation helps a lot in image recognition. Image segmentation extracts the key features of an image.By observing the model segmentation in Fig. 10, it can be understood that it is the human face of a girl who is laughing. It also separates the foreground from the background, which makes it easier for the blackbox model to understand the real subject of the sample image. To give a correct prediction, understanding the subject matter of the image is a must. Fig. 10 indicates that the proposed model has the proper training to identify the subject matter of an image and can separate foreground from background. The key defining

features of the sample image are isolated using image segmentation. The localization of the facial features made the sample image easy to understand. While processing the sample image, the image was converted into a collection of regions with masking. The common factor in all the small segmented parts is that, in Fig. 10, they all have similar attributes. Performing image segmentation is difficult, but with the LIME algorithm, it is possible to get a reliable outcome.

Figure 9: Sample image of human face

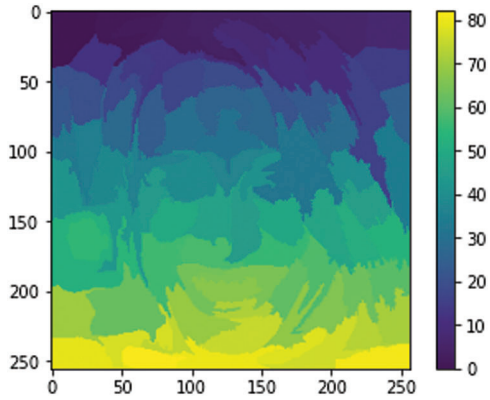


Figure 10: Model segmentation

Next, in Fig. 11, it is very clear that the model has a proper understanding of the sample face. With the help of 3D masking in the segmentation algorithm, the cheeks, teeth, and eyes are all separated. This 3D image boundary makes the model more reliable and trustworthy. A scikit-learn image segmentation function was carried out for the 3D boundary. In Fig. 11, the 3D masking creates a clear label around the sample image to create a 3D depth of field around the face. This represents the model's awareness that the intended target is correct. A model boundary is a visual representation of the model's understanding of an unseen image. A properly defined 3D boundary represents good model training that will perform well on unseen datasets. In practice, the DL model will face problems where defining boundaries will result in correct predictions most of the time. XAI makes it interpretable and informational.

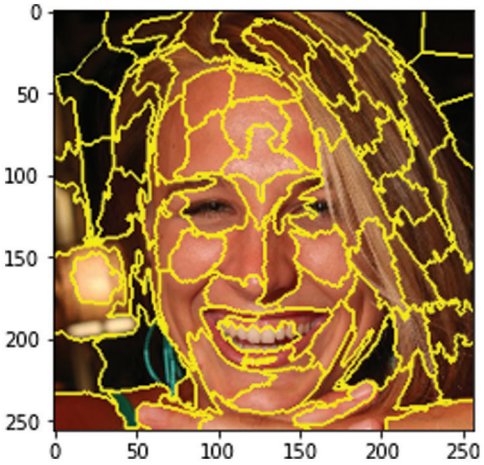


Figure 11: Model boundaries

The LIME algorithm is model-agnostic and can explain both classification and regression models. The proposed model uses the LIME image explainer. The image had to be a 3D NumPy array to maintain the format of the explainer; otherwise, the explainer could not detect the image. It describes image prediction by starting with a sample of 0 and inverting the mean-centering and scaling methods. Our model samples the training distribution while it solves a classification issue, and when the value is the same, it creates a binary feature that is 1. The instance from the LIME image generates neighborhood data after the explainer is set. Explanations can be retrieved from the model after learning the weight of the linear models locally. The highest weight of the prediction probability considered for that particular sample image is shown in the top labels. Finally, the model can explain any prediction's significant weight. For the above sample image in Fig. 12, the model puts more emphasis on the red-highlighted portion of the image, which matches the sample's intended target. The red-highlighted portion was the face of the person, so the proposed model prediction is based on the correct subject. Deepfake images of a human face usually change key features like human eyes, nose, and mouth, and the proposed model shows accountability in these areas of the sample images in Fig. 12 by marking these areas red while showing image temperature. Sometimes the affected areas can be defined vividly, other times the model cannot put any image temperature over that area, like in Fig. 13 in the hair regions.

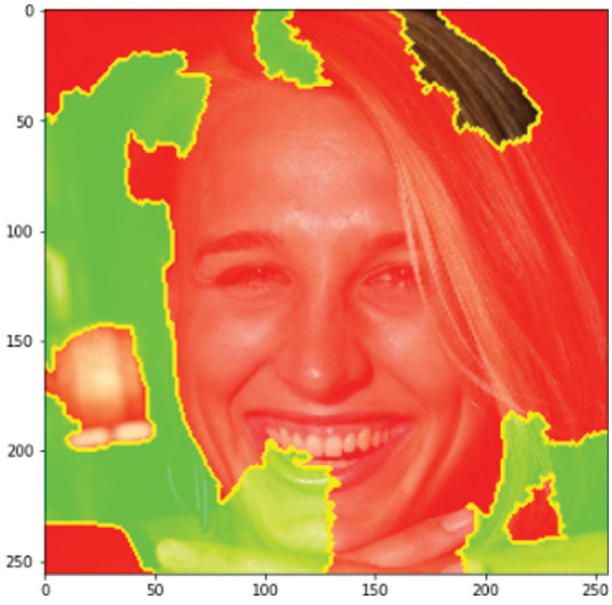


Figure 12: Image temperature

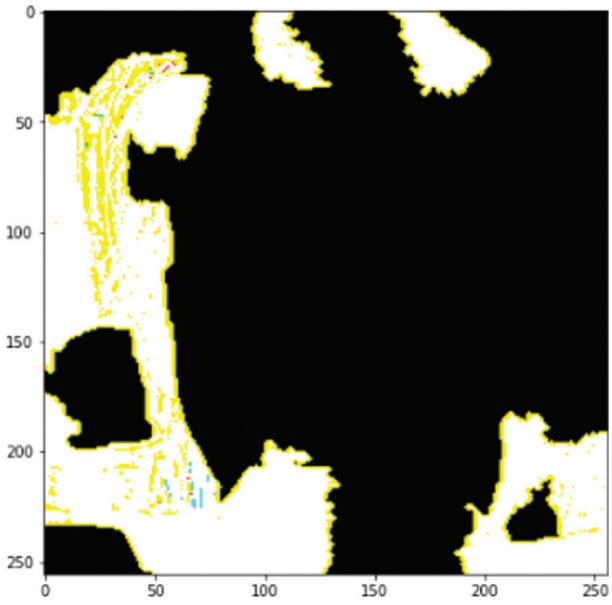


Figure 13: Image temperature (positive only)

Again, in Fig. 14, only the part that is only positive remains false in colored black. This makes looking at the sample much easier. By focusing on the highlighted area, it is easier to identify deepfake images. It can be easily understood that the factors for the prediction of the model are in the black zone. The sample image had the face in the center, which is marked black using XAI. The black box model gives a prediction of an image using only the positive areas of a sample image. That is why a good model needs to be able to identify the weighted zones. With the help of XAI, the proposed model is proved to have a strong understanding of unseen sample images.

The accuracy of the proposed model (InceptionResNetV2) was 99.87%. The training loss in the final epoch for the model was almost nil with the most explicit black-box explanation. Moreover, InceptionResNetV2 achieved the highest validation accuracy. Finally, XAI gives a clear understanding and trust in deep neural models. XAI gives the user a hint behind the model prediction, so this approach is more flexible and reliable. Users are more likely to accept the prediction of a model with XAI than a blackbox model.

3.4 Model Comparison

The best trained models used in this paper were compared to those in the above-referred papers. The accuracy is given in Tab. 4. With the help of Explainable AI (XAI), the accuracy of all our trained models is efficient for detecting deep-fake images. In that study [23], the GAN system was used to detect deepfake images with 90.22% accuracy. The accuracy of the study [23] is not high enough to be a stable model. The proposed model in this paper has much higher accuracy than the model evaluation using XAI, which makes the proposed model in this paper more robust. Besides accuracy, the study [23] does not provide any concrete evaluation. The journal paper [25] used the VGG convolutional method to detect deepfake face images. Again, the accuracy of the journal paper [25] is on the lower side. Also, the image size of the dataset was only 128 × 128 px, whereas the method mentioned above uses a 224 × 224 px image. In the research paper [27], a transfer learning method based on the GAN architecture was used to detect deepfake images. A few methods of transfer learning, like VGG16 and MobileNet, were carried out in the paper [27], but every method gave average accuracy. The method mentioned in this

paper has low validation loss and high validation accuracy. In the discourse [28], the transfer learning method of the Convolutional Neural Network (CNN) architecture was chosen to carry out the research, which achieved an accuracy of 91.56%. But the learning rate used in this research is 0.00001, which is very low and makes it harder for any model to learn new information in a single epoch. Also, the batch size is 256, whereas the standard batch size is 64. In the study [29], recurrent neural networks were performed to detect deepfake images with an accuracy achieved of 97.83%, which is good, but the model proposed in this paper still has the highest accuracy. These are the conditions under which the proposed method can help detect deepfake images.

Table 4: Result comparison with previous works

Paper	Accuracy	This paper accuracy
In study [23]	90.22%	99.87%
In study [25]	85.70%	
In study [27]	90.20%	
In study [28]	91.56%	
In study [29]	97.83%	

4 Conclusion

The main objectives of this project are to classify real and fake images, show the accuracy of different Convolutional Neural Network (CNN) algorithms, and lastly, verify and explain the model using Explainable AI (XAI) so that it becomes clear which algorithm is more effective in detecting deepfake. Our research will shed light on the use of and help in day-to-day life. The use of XAI in deepfake image detection is novel and is exclusively present in this research paper. The proposed system provides 99.87% accuracy in detecting deepfake images from real images. The results indicate that the proposed method is very robust at detecting deepfake images and is also reliable and trustworthy because of the verification and integration from XAI. The Local Interpretable Model-Agnostic Explanations (LIME) algorithm is also utilized in this method to describe which component of the image from the dataset caused the model to create specific classifications, ensuring the model's validity and dependability. This will help future researchers select CNN algorithms to differentiate between deepfakes and their effects. It will help people be aware of deepfake content and will also give relative information about CNNs and their working procedures. People can learn about choosing and training different datasets. If this system were available on the market, it would save many people the hassle of dealing with fake content and fake information. Advanced research will significantly improve the general state of this system and its explanation in the future. Furthermore, the project can be improved by examining the model with updated transfer learning methods from its new versions. Presently, XAI is limited to images, but in the future, video samples could be validated by XAI. Also, the model could be directed to focus on specific features of a face to detect deep fake images. These minor updates can make this model more versatile. It will yield behavioral conclusions as well as important insights into the operations of deep networks. Researchers may have to look at medical images to implement and get the best outcome from deepfake detection for further medical analysis. This will increase trust in deep learning systems while also allowing for better understanding and enhancement of system behavior.

Acknowledgement: The authors would like to thank Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R193), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors are thankful for the support from Taif University Researchers Supporting Project (TURSP-2020/26), Taif University, Taif, Saudi Arabia.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R193), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Taif University Researchers Supporting Project (TURSP-2020/26), Taif University, Taif, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] U. Rahul, M. Ragul, K. R. Vignesh and K. Tejeswini, "Deepfake video forensics based on transfer learning," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 5069–5073, 2020.

[2] FaceApp - AI face editor," Faceapp.com. [Online]. Available: https://www.faceapp.com.

[3] G. Benjamin, "Deepfake videos could destroy trust in society–here's how to restore it," The Conversation, 2019. [Online]. Available: https://theconversation.com/deepfake-videos-could-destroy-trust-in-society-heres-how-to-restore-it-110999.

[4] R. Scammell, "Deepfakes: AI video tool could make fake news easier to create and harder to spot," Verdict, 2018. [Online]. Available: https://www.verdict.co.uk/deepfakes-ai-video-tool-fake-news.

[5] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, pp. 1–6, 2018.

[6] "Deepfake laws risk creating more problems than they solve," Regulatory Transparency Project, 2021. [Online]. Available: https://regproject.org/paper/deepfake-laws-risk-creating-more-problems-than-they-solve.

[7] J. Damiani, "A voice deepfake was used to scam a CEO out of $243,000," Forbes, 2019. [Online]. Available: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=66e1686a2241.

[8] E. J. Dickson, "Deepfake porn is still a threat, particularly for K-Pop stars," Rolling Stone, 2019. [Online]. Available: https://www.rollingstone.com/culture/culture-news/deepfakes-nonconsensual-porn-study-kpop-895605.

[9] H. S. Shad, M. M. Rizvee, N. T. Roza, S. M. A. Hoq, M. M. Khanet *et al.,* "Comparative analysis of deepfake image detection method using convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2021, no. 3111676, pp. 1–18, 2021.

[10] "Report: Number of deepfakes double every six months," CyberNews, 2021. [Online]. Available: https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months.

[11] J. E. Solsman, "Deepfakes' threat to the 2020 US election isn't what you'd think," CNET, 2020. [Online]. Available: https://www.cnet.com/features/deepfakes-threat-2020-us-election-isnt-what-youd-think.

[12] Y. L. Cun, Y. Bengio and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big Data*, vol. 5, no. 3, pp. 246–255, 2017.

[14] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[15] "Explainable AI," Ibm.com. [Online]. Available: https://www.ibm.com/watson/explainable-ai.

[16] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He and X. Chen, "Tbe-net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 1, pp. 1–13, 2021.

[17] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 1, no. 3, pp. 1–16, 2021.

[18] N. Baghel, U. Verma and K. K. Nagwanshi, "WBCS-net: Type identification of white blood cells using convolutional neural network," *Multimedia Tools and Applications*, vol. 1, no. 1, pp. 1–10, 2021.

[19] K. N. Alam, M. S. Khan, A. R. Dhruba, M. M. Khan, J. F. Al-Amri *et al.,* "Deep learning-based sentiment analysis of COVID-19 vaccination responses fromtwitter data," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 4321131, pp. 1–15, 2021.

[20] M. H. M. Khan, N. B. Jahangeer, W. Dullull, S. Nathire, X. Gao *et al.,* "Class Classification of Breast Cancer Abnormalities Using Deep Convolutional Neural Network (CNN)," *PLOS ONE*, vol. 16, no. 8, pp. 1–15, 2021.

[21] J. Gupta, S. Pathak and G. Kumar, "Bare skin image classification using convolution neural network," *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 1, pp. 138–145, 2022.

[22] K. N. Alam and M. M. Khan, "CNN based COVID-19 prediction from chest x-ray images," in *Proc. 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conf. (UEMCON)*, New York, USA, pp. 486–492, 2021.

[23] L. Guarnera, O. Giudice and S. Battiato, "DeepFake detection by analyzing convolutional traces," in *Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, pp. 2841–2850, 2020.

[24] S. Saha, "A comprehensive guide to convolutional neural networks—the ELI5 way," Towards Data Science, 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[25] X. Chang, J. Wu, T. Yang and G. Feng, "DeepFake face image detection based on improved VGG convolutional neural network," in *Proc.39th Chinese Control Conf. (CCC)*, Shenyang, China, pp. 7252–7256, 2020.

[26] C. C. Hsu, Y. X. Zhuang and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Applied Sciences*, vol. 10, no. 1, pp. 370, 2020.

[27] M. Patel, A. Gupta, S. Tanwar and M. S. Obaidat, "Trans-DF: A transfer learning-based end-to-end deepfake detector," in *Proc. IEEE 5th Int. Conf. on Computing Communication and Automation (ICCCA)*, Greater Noida, India, pp. 796–801, 2020.

[28] S. Suratkar, E. Johnson, K. Variyambat, M. Panchal and F. Kazi, "Employing transfer-learning based CNN architectures to enhance thegeneralizability of deepfake detection," in *Proc. 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. 1–9, 2020.

[29] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson *et al.,* "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[30] C. Rudin and J. Radin, "Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition," *Harvard Data Science Review*, vol. 1, no. 2, pp. 1–10, 2019.

[31] Understanding black-box ML models with explainable AI," Dynatrace.com. *Available:* https://engineering.dynatrace.com/blog/understanding-black-box-ml-models-with-explainable-ai.

[32] J. Petch, S. Di and W. Nelson, "Opening the black box: The promise and limitations of explainable machine learning in cardiology," *Canadian Journal of Cardiology*, vol. 38, no. 2, pp. 204–213, 2022.

[33] C. Schorr, P. Goodarzi, F. Chen and T. Dahmen, "Neuroscope: An explainable AI toolbox for semantic segmentation and image classification of convolutional neural nets," *Appl Sci. (Basel)*, vol. 11, no. 5, pp. 2199, 2021.

[34] F. Stieler, F. Rabe and B. Bauer, "Towards domain-specific explainable AI: Model interpretation of a skin image classifier using ahuman approach," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, pp. 1802–1809, 2021.

[35] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should i trust you?:Explaining the predictions of any classifier," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 1114–1135, 2016.

[36] 140k Real and Fake Faces," Kaggle. [Online]. Available: https://www.kaggle.com/xhlulu/140k-real-and-fake-faces.

[37] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 4396–4405, 2019.

[38] D. Johnson, "What is a deepfake? Everything you need to know about the AI-powered fake media," Business Insider, 2021. [Online]. Available: https://www.businessinsider.com/what-is-deepfake.