DEEPFAKE DETECTION USING CONVOLUTION NEURAL NETWORK

Supported by

Agu, Edward .O.; Dennis, Samuel Toochukwu
Computer Science Department, Federal University Wukari, Taraba, Nigeria
Corresponding author: aguedwardo@gmail.com

Received: February 11, 2023 **Accepted:** April 15, 2023

Abstract:

Over the past few decades, artificial intelligence (AI) has been developing rapidly. It has applications in various fields for countless purposes. But not all AI products have a positive impact on society. Sometimes, technologies created for good reasons are later abused by criminals. An example of such technology is known as Deepfake. Deepfake is a technology that allows one to replace a face or change the mouth movement of a person and facial expressions to make them say whatever they wish using any of the deepfake manipulation techniques which include NeuralTextures, Face2Face, Deepfakes, and FaceSwap. Despite its usefulness, if used maliciously, it can severely impact society, for instance, spreading fake news and cyberbullying, among others. Using Deepfakes manipulation techniques, this study, therefore, aim to address this problem by proposing a model that analyses the frames of a video using a deep learning approach to detect the forged areas in the video and to deploy the trained model using Flask (a framework used for the deployment of web applications in python). The proposed model uses an EfficientNet B6 classifier to train a neural model to detect deepfake images via a FaceForensics++ dataset. The trained model was able to classify videos with an accuracy 90%. To validate the model performance, the precision, recall and f1-score was utilized with a result of 98%, 81% and 89% respectively for the class of fake images and a result of 84%, 98% and 91% for the class of real images. The motive of the study was to improve learning by this model.

Keyword: Deepfake, FaceSwap, Detection, Naturaltextures, facial_expression

Introduction.

For the past decades deepfake technology has been used for notorious pranks, committing online fraud, influencing the general public viewpoint, and embarrassing political officials which need to be checked using sophisticated mitigation tools (Agu *et. al*, 2017; Francisca*et. al*, 2015). Furthermore, the technology also poses a great threat to biometric facial recognition technology by utilizing the Generative Adversarial Network (GAN). A Generative Adversarial Network (GAN) is an analytical technology that can produce false positives and false negatives through fake videos and pictures. DeepFakes emerges as one of the latest manifestations of GAN, which create exceptionally well counterfeit pictures and videos which are quite hard to differentiate from the originals.

In recent years, technologies to manipulate facial videos have reached thepoint where it might not be possible for a human to detect that the video is manipulated. (Emil Johansson, 2020). Deepfake videos might pose a threat to this assurance. A Deepfake algorithm can create real-time video manipulations in which one could insert a person into a video by pasting their face onto thatof another personor change a person's mouth movementand facial expressions to make them say whatever they wish. This replacement can be done with FaceSwap, face2face, NeuralTextures, Deepfakes, and facial reenactment, or by generating a whole new video starring the target individual. It is not hard to imagine what harm these kinds of videos could cause if createdwith malicious intent; the resulting videos vary from fake pornographic videos to fake political speeches.

In a culture that is rife with misinformation and disinformation, it can be easy for people to be duped into believing they are reading or seeing something that has no base. Deepfake videos have added to this confusion, sometimes presenting content that is meant to deceive the viewer or to drastically misrepresent the person in the video. With the advent of deepfakes, viewers now need to question whether what they are seeing in a video is real or not. Fake videos have been made of politicians endorsing views contrary to their own, public figures confessing to wrongdoings, and women engaging in sexual activities they never engaged in. Some of these videos are deepfakes, due to their low-quality visual effects, unusual contextual setting, or the explicit acknowledgment that they are deepfakes. But many others are nearly impossible to distinguish from a real video and are not labeled as fakes. (Dunn., 2021).

The detection of deepfake videos can be considered as a binary classification problem because every imageis either 'real' or 'fake'. This research aims to solve the problem of deep face impersonation by implementing a EfficientNet model that analyses the frames of the videos using a deep learning approach to detect inconsistencies in facial features, introduced in the videos while creating the frames. It uses a dataset called "FaceForensics++".

Artificial Neural Networks (ANN) are algorithms based on brain function and are used to model complicated patterns and forecast issues. The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain's Biological Neural Networks. The development of ANN was the result of an attempt to replicate the workings of the human brain. The workings of ANN are extremely like the brains of biological neural networks, although they are not identical. ANN algorithm accepts only numeric and structured data. (Gourav, 2021).

Before 2006, convolutional networks (and their predecessors) were the only deep networks that could be trained successfully(Goodfellow, Bengio, and Courville, 2016). In 2006, the Deep Belief Network (DBN) changed this. Subsequently, the term 'deep learning' was first introduced in 2006 (Caterini and Chang., 2018). Since then, neural networks have come a long way. Although neural networks might seem magical and futuristic, getting from an idea of something that could be solved by a neural network to a working neural network that can perform the task in mind can take a while. Data needs to be collected (and sometimes labeled) and preprocessed, and in the meantime, an algorithm must be designed. Once that is done, the algorithm must be trained. Training is not always stable; the process might have to be repeated several times until the results are satisfactory. Moreover, a network that performs well on a training dataset need not necessarily perform well on unseen data. (Leeuwen, 2020).

ANN Activation function defines the output of input or set of inputs or in other terms defines node of the output of node that is given in inputs. They decide to activate or deactivate neurons to get the desired output. It also performs a non-linear transformation on the input to get better results on a complex neural network.

The activation function also helps to normalize the output of any input in the range between 1 to -1 or 0 to 1. The activation function must be efficient, and it should reduce the computation time because the neural network is sometimes trained on millions of points (Leeuwen, 2020). Binary step, linear, and non-linear is the types of activation.

A convolutional neural network (CNN) is a particular implementation of a neural network used in machine learning that exclusively processes array data such as images and is thus frequently used in machine learning applications targeted at medical images. A convolutional neural network consists of the following three components: input (image), feature extraction, and non-linear activation unit. The kernel can be understood as a small 2-D matrix that is used for the case of establishing a relationship of the center pixel with respect to its neighboring pixels. (Shubh Saxena- August 2019).

In deepfake images, EfficientNet uses a technique called compound coefficient to scale up models in a simple but effective manner. Instead of randomly scaling up width, depth or resolution, compound scaling uniformly scales each dimension with a certain fixed set of scaling coefficients. Using the scaling method and AutoML, the authors of efficient developed seven models of various dimensions, which surpassed the state-of-the-art accuracy of most convolutional neural networks, and with much better efficiency(Sarkar, 2021).

EfficientNet is based on the baseline network developed by the neural architecture search using the AutoML MNAS framework. The network is fine-tuned for obtaining maximum accuracy but is also penalized if the network is very computationally heavy. There are seven different EfficientNet models of which EfficientNet B7 is the biggest obtained state-of-the-art performance on the ImageNet and the CIFAR-100 datasets. It obtained around 84.4% top-1/and 97.3% top-5 accuracy on ImageNet. Also, the model size was 8.4 times smaller and 6.1 times faster than the previous best CNN model. It obtained 91.7% accuracy on the CIFAR-100 dataset and 98.8% accuracy on the Flowers dataset. (Sarkar, 2021).

MTCNN is a neural network that detects faces and facial landmarks on images. It was published in 2016 by Zhang et al. MTCNN is one of the most popular and most accurate face detection tools today. It consists of three neural networks connected in a cascade. (Adamczyk., 2021).

Deepfakes have become popular due to the qualityof tampered videos and the easy-to-use ability oftheir applications to a wide range of users with variouscomputer skills from professional to novice. These applications are mostly developed based on deep learningtechniques. Deep learning is well known for its capabilityof representing complex and high-dimensional data. Onevariant of the deep networks with that capability isdeep autoencoders, which have been widely appliedfor dimensionality reduction and image compression(Santha, 2020).

Flask framework used deepfake was originally designed and developed by Armin Ronacher as an April Fool's Day joke in 2010. Despite the origin as a joke, the Flask framework became wildly popular as an alternative to Django projects with their monolithic structure and dependencies. (www.python.com/flask.html).

Review of Related Literatures:

According to Guera and Edward, (2018) research on Convolution Neural Network, Recurrent NeuralNetwork tried to evaluate the method against alarge set of DeepFake videos collected from multiple videowebsites. They propose a temporal recognition pipeline to automatically detect fake videos. This system works only with a large dataset.

Bayar and Stamm, (2016) demonstrated an 8 layers CNN-based network: a constrained convolutional layer, 2 additional convolutional layers with 2 Max-pooling layers, and 3 fully connected layers. Their method has achieved 86.10% accuracy on easily compressed videos and 73.63% accuracy on Strongly compressed videos.

Li *et al.*, (2021) in their research based on Convolution Neural Networks and Recursive Neuralnetworks tried to create a new system thatexposes fake faces based on eye blinking, that has beengenerated using Neural Networks. Therefore, in hispaper, he aimed at analyzing the eye blinking in thevideos, which is a psychological signal that is not well presentedin the synthesized fake videos. He also used VGG16 as a CNNmodel to distinguish eye states.

Rahmouni *et al.*, (2017) trained a CNN with a custom pooling layer to optimize the feature extraction algorithms. By local estimates of class probabilities to predict the label of an image, they achieved 88.5% testing accuracy on easy compressed videos and 61.5% testing accuracy on Strongly compressed videos on the FaceForensics dataset.

Agarwal *et al.*, (2019) in their paper on Protecting the World Leaders Against Deep Fakes, employ the OpenFace2 toolkit to classify various facial features such as the mouth, nose, and lip. The partial accuracy is over 90%.

Hasan and Salah, (2019) in Combating deepfakeVideos Using Blockchain and Smart Contracts” wrote apaper on Blockchain Technology and Artificial Intelligence.The author proposes a blockchain-based system for deepfakevideos. The system provides a trusted way for secondary artists to requestpermission from the original artist to copy and edit videos.

Rana *et al.*,(2021) wrote a paper on Convolution Neural Network, Dual tree complexwavelet transform (DT DCT), Depth image-based rendering(DIBR), Multiview video plus depth (MVD), 3D highlyefficient-video-coding(3D-HEVC). In this, they tried todetect methods to differentiate fake 3D video and real 3D videousing CNN. The research work is based on identifying the real and fake 3D andpre-filtration is done using the dual-tree complex wavelettransform to emerge the edge and vertical and horizontalparallax characteristics of real and fake 3D videos. High-resolution video sequences are used for training. Theyimplemented CNN architecture for the proposed scheme.

Abdali *et al.,* (August 2021) in their research on deepfake representation with multilinear regression did a similar Deepfake video classification model using the FaceForensics++ dataset. He used SVM classification with an accuracy of around 82 %.

Raghavendra *et al.,* (2017) used two fully connected CNN (VGG19 and AlexNet) to detect the feature, followed by a probabilistic collaborative Representation Classifier (P-CRC) to detect the morphed images. They achieved 93.5% accuracy on 10 easy compressed videos and 82.13% accuracy on Strongly compressed videos on FaceForensics dataset.

According to Karandikar (2020) research on Deepfake Video Detection Using Convolutional Neural Network. The author used VGG16 classification to detect Deepfakes videos with an accuracy of 70%.

Problem and Proposed Solution Analysis

Analysis of the Existing System

Table 3.1. Existing system

Author	Published	Method	Dataset	Result
Raghavendra *et al.*	2021	VGG16 and DenseNet	FaceForensics	95% and 94%
Aarti Karandikar	2020	VGG16	Celeb-DF	70%
Abdali *et al.*	2021	SVM (Support Vector Machine	FaceForensics++	82%

Analysis of the Proposed System

The steps taken into cognizance after reading the fake and real video incorporates data preprocessing steps that extract faces from both the fake and real video using MTCNN (MTCNN (Multi-Task Cascaded Convolutional Neural Networks**).** After the extraction step, the images extracted are converted into a vectorized array of numerical representations. Hence, the vectorized image values are fed to the EfficientNet before a fully connected neural network model is developed and validated using the test data and also using some evaluation metrics such as the confusion metrics, precision score, F1-score, and accuracy scores. The figure below shows the steps utilized to train and validate the model after reading the FaceForensics++ dataset.

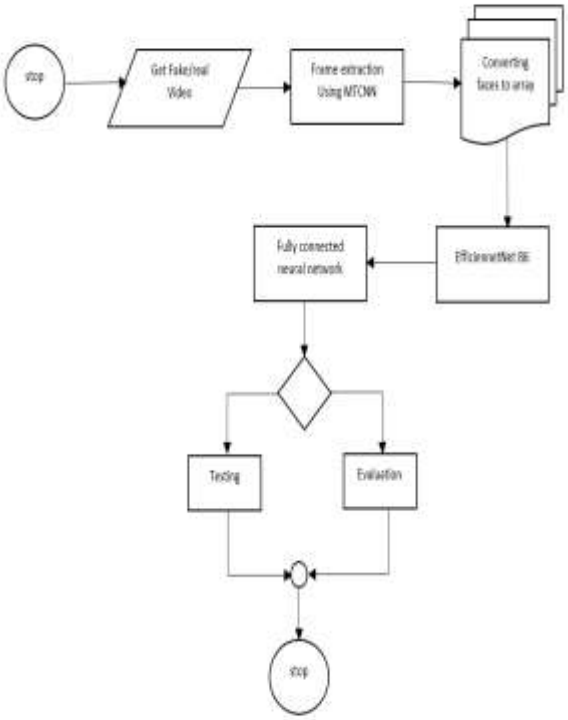

stop
Get Fake/real Video
Frame extraction Using MTCNN
Converting faces to array
EfficienetNet B6
Fully connected neural network
Testing
Evaluation
stop

Figure 3.1. Process flow for deepfake detection

Dataset

FaceForensics++ is a forensics dataset consisting of 1000 original YouTube videos that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures containing 1000 videos. In this paper, an experiment is carried out on the Deepfakes subset for fake videos. The dataset was downloaded from (www.kaggle.com). The size of the entire data is about 10 gigabytes. The videos have a wide range of facial expressions because they are about TV reporters and journalists of various sexes, ages, and races. In short, FaceForensics++ contains 5000 both real and fake videos as shown below.

Table 3.2. Dataset Description

Manipulation method	Number of videos
Deepfakes	1000
Face2Face	1000
FaceSwap	1000
NeuralTextures	1000
Real videos from YouTube	1000

Model Learning Processes

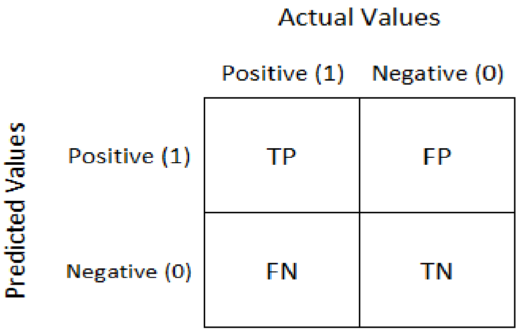
Dataset (FaceForensics++) consists of 1000 real and fake videos which are preprocessed. This involves extraction of frames using OpenCV (a python library) and face extraction and alignment using MTCNN. MTCNN is a neural network that detects faces and facial landmarks on images. It was published in 2016 by Zhang et al. MTCNN is one of the most popular and accurate face detection tools today. It consists of three neural networks connected in a cascade. (Adamczyk., 2021).The proposed model targets faults induced during deepfake creation around the face outline. Thus, face extraction will extract the area that needs to be processed. Face alignment is used to account for different head positions that the target person may have in the deepfake video. After pre-processing, the faces are converted into an n-dimensional array of size (320,320,3)where 320 is the height and width whereas 3 represents thenumber of channels. Further, the inputs are flattened to make themof the size (320*320*3,1) to give as an input to the first layerof the neural network. Thus, the data is made ready to be givenas input. The model uses the pre-processed frameset from the original dataset and implements transfer learning on a fine-tuned EfficientNet B6 model for the detection of Deepfakes.

The proposed method consists of the EfficientNet B6 model (trained on ImageNet dataset by Google) as its base, dropout, and a custom three-node dense layer. The third node in the last dense layer in the architecture proposed is used for two final classes (real and fake). Further, dropout is added to reduce overfitting and better optimization of weights. The dropout layer will randomly send some nodes as off from the previous layer at every epoch. This will lead to better training as some randomness is induced by this layer while updating weights. The hyperparameter tuning is done and thenumber of layers, activation function, optimizers, and learningrate. Feature extraction is done using the convolution operationusing 3x3 filters. Adam Optimizer gives the best learning for this scenario. A learning rate of around 0.001 is used for successful feature extraction and training.

Performance Evaluation Metrics

Confusion Matrix

A confusion matrix is a performance measurement for a machine learning classification problem where the output can be two or more classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives a holistic view of how well the classification model is performing and what kinds of errors it is making. It is a table with four different combinations of predicted and actual values as shown below. (Sarang, 2021).



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.2. Confusion matrix table

- **True Positive (TP):**The actual value was positive, and the model predicted a positive value
- **True Negative (TN):**The actual value was negative, and the model predicted a negative value
- **False Positive (FP):** The actual value was negative, and the model predicted a positive value. Also known as **Type 1 error.**
- **False Negative (FN):** The actual value was positive, and the model predicted a negative value. Also known as **Type 2 error.**

The accuracy of a classification model is calculated using the formula below.

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

Recall

Defines the number of actual positive values that the model was able to predict correctly. A mathematical representation for the recall metric can be expressed as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision

It defines how many of the correctly predicted values turned out to be positive. It is calculated using the formula below.

$$\text{Precision} = \frac{TP}{TP + FP}$$

In a situation where there is no clear distinction between whether Precision is more important or Recall, they are combined to what is known as **F1-Score**. It is calculated using the formula below.

$$\text{F1-Score} = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

Methodology Adopted

This research proposes a method to train the classifier based on video frames as input. The frames are passed through face extraction and alignment fragment and then passed to the classifier for training using the EfficientNet B6 model as it is the best model so far. The figures below show the steps that will be followed to train the model.

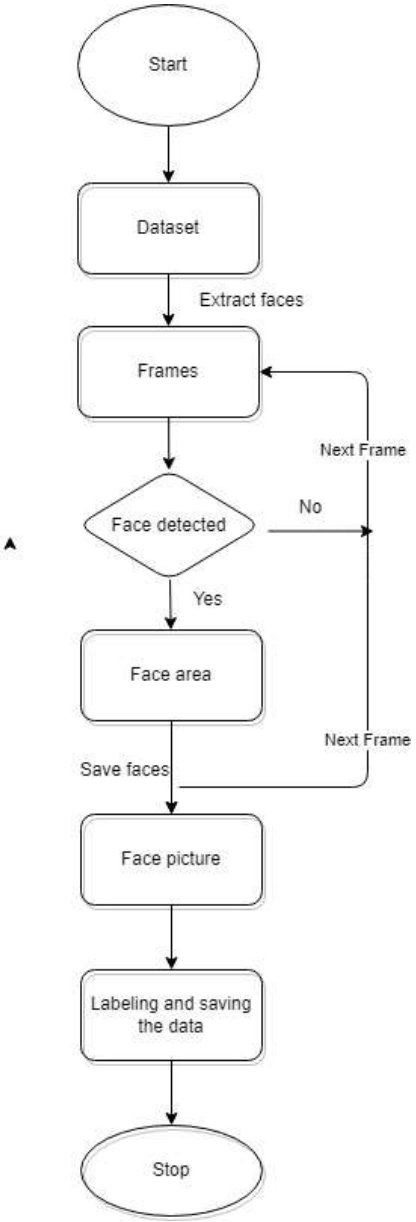


Figure 4.1. Preprocessing

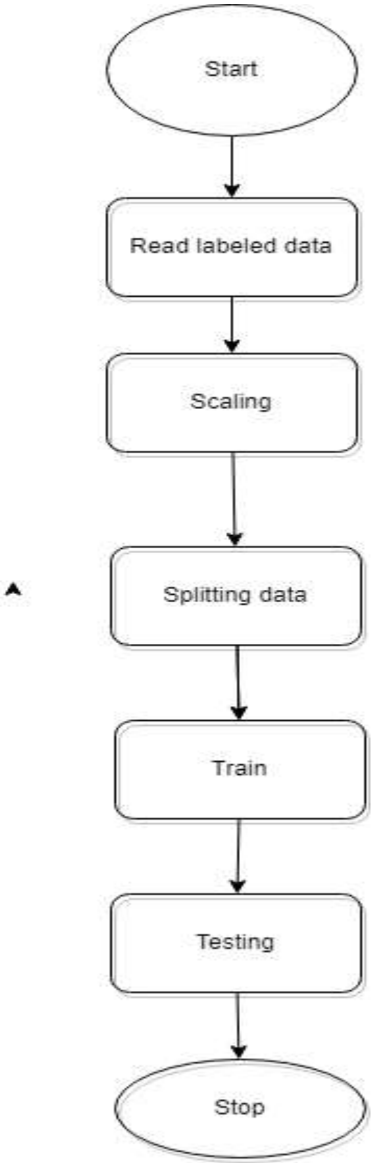


Figure 4.2. Deep learning model

Implementation

Extract Video Frames and Save to Images

As stated by the authors above, all videos have a constant frame rate of 30 fps (frames per second). So, 7 frames are extracted from each video and saved as images. To be more specific, 1 frame is extracted in every 30 frames which means 7 frames are extracted as images in one video. In this case, different facial expressions can be captured in a single video. Now there are 1000 Deepfake and original videos. After the extraction, there are 14000 images which include 7000 "fake" images and 7000 "real" images, cv2, and imageio were used to capture features from the given videos. Examples of both fake and real frames are shown in the figures below.

Figure 5.1. Frames extracted from deepfake videos

Figure 5.2. Frames extracted from real videos

Extract Face from Saved Frames Using MTCNN

MTCNN (Multi-Task Cascaded Convolutional Neural Networks) is a type of neural network that recognizes faces and facial landmarks in images and further provides the exact pixel positions of the precise nose, mouth, left eye, right eye, and face boundary.

If one image is read, the MTCNN model returns three indexes: box, confidence, and key points. Some face shapes are unclear so, one more condition was added to set confidence>0.9 in the facial extraction function, which means that if MTCNN does not have 90% confidence to identify a face in an image, it should skip it because it is an outlier.

Finally, 6985 out of 7000 facial images in the fake set and 6999 out of 7000 facial images in the real set were captured. The capture rate of Deepfake images is 99.77%, and the capture rate of original images is 99.9%. The image size was unified to 320 x 320 x 3 using the PIL package. In conclusion, now there are 13985 observations, which include 6984 fake images and 6999 real images. Examples of faces are shown in Figures 5.3 and 5.4.

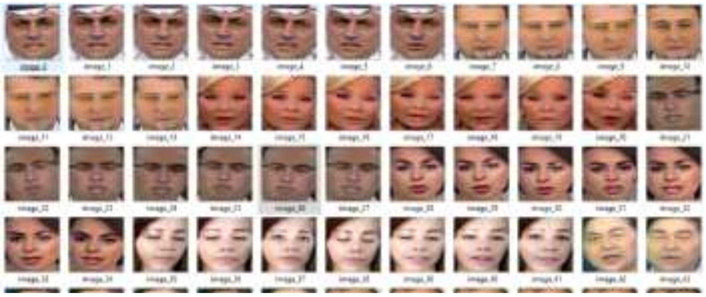


Figure 5.3. Faces extracted from deepfake frames

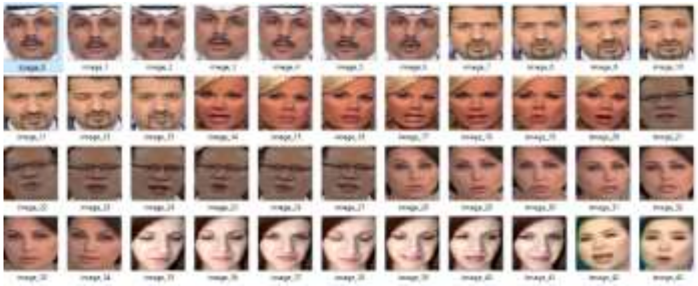


Figure 5.4. Faces extracted from real frames

After the extraction, the faces were saved in an array labeling the real faces as ‘1’ and fake faces as ‘0’.

Data Visualization

There are 13984 observations, which include 6985 fake rows and 6999 real rows.

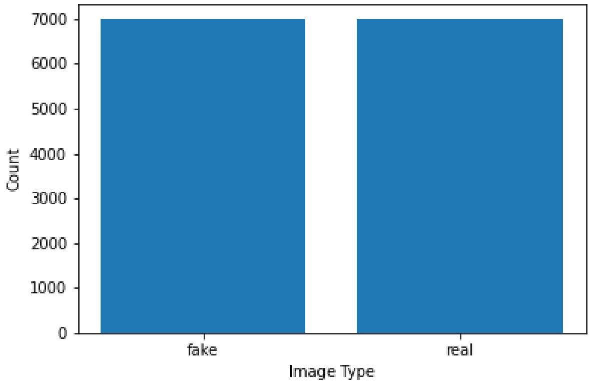


Figure 5.5. Data visualization

Training

80% of the data was used to train the model for 5 epochs using the accuracy metrics to validate the model during the training. The trained model was tested on the remaining data. An accuracy of 90% was achieved. Using Flask framework, the trained model was deployed for real-time testing.

Results

After training and testing the model, accuracy of around 90% was achieved based on the features learned by image analysis. The confusion matrix and classification report of the model are shown below.

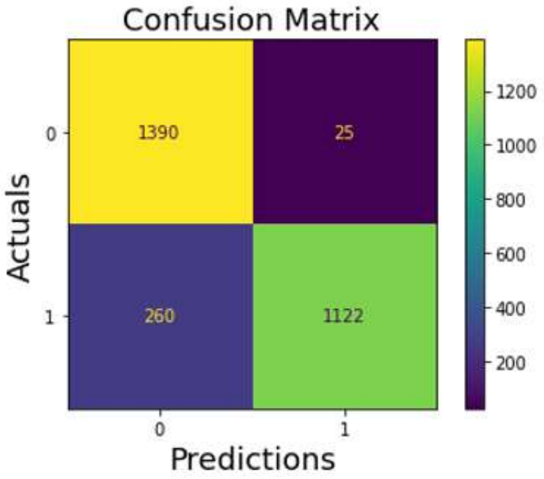
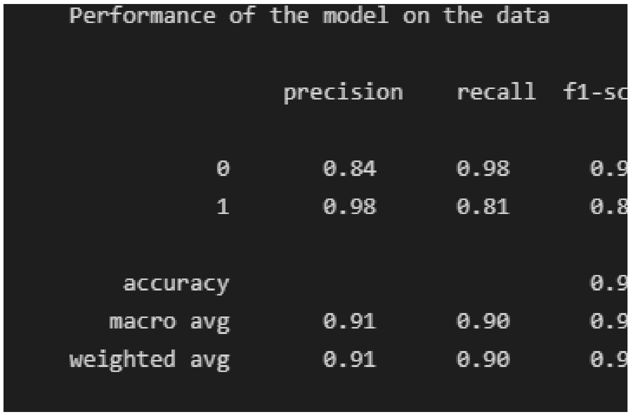


Figure 6.1. Confusion matrix of the model



```
    Performance of the model on the data

                   precision    recall  f1-sc

             0        0.84        0.98      0.9
             1        0.98        0.81      0.8

      accuracy                              0.9
     macro avg        0.91        0.90      0.9
  weighted avg        0.91        0.90      0.9
```

Figure 6.1. Classification report

Conclusion

Detecting whether a video is manipulated nowadays is important given the significant impact of videos in everyday life and online communication. Consequently, this study focused on training a CNN model in hybridization with an EfficientNet B6 model and thus, deploying it for real-life testing by the public to check the originality of fake videos generated by deepfake manipulation technology.The analytical result from the trained model obtained an accuracy of 90%, to validate the model accuracy, the precision, recall, and f1-score were utilized with a result of 98%, 81%, and 89% respectively for the detected fake images and a result of 84%, 98% and 91% for the real images.

Recommendation

There are several ways to improve the proposed model, such as increasing the number of epochs for better performance and accurate results on the dataset. Training the feature extraction network and some polishing of hyperparameters would most likely increase the model's accuracy drastically. To increase the utility of the model, more manipulation methods could be considered.

References

Abdali, S. (2021, August 15). *Deepfake Representation with Multilinear Regression*. arXiv.Org. Retrieved May 14, 2022, from https://arxiv.org/abs/2108.06702

Agu, E. O., Ejiofor V. E.,Moses T., (2017) A Hybrid Model For Remote Dynamic Data Auditing (RDDA) On Cloud Computing. Journal of Computer Science and Application (JCSA). Vol. 24, No. 1, pp 96-116.

Burns, E., & Brush, K. (2021, March 29). *deep learning*. SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019). Energy compaction-based image compression using convolutional autoencoder. IEEE Transactions on Multimedia. DOI: 10.1109/TMM.2019.2938345.

Chorowski, J., Weiss, R. J., Bengio, S., and Oord, A. V. D. (2019). Unsupervised speech representation learning using wavelet autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 27(12),2041-2053

D. Guera and E. J. Delp, Deepfake Video Detection Using Recurrent Neural Networks, 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp.1-6.

Davis E. King. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10:1755–1758, 2009.

Dunn, S. (2021, March 3). *Women, Not Politicians, Are Targeted Most Often by Deepfake Videos*. Centre for International Governance Innovation. Retrieved May 14, 2022, from https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/

Education, I. C. (2021, August 3). *Neural Networks*. Neural Network. https://www.ibm.com/cloud/learn/neural-networks

Face Recognition with FaceNet and MTCNN. (2020, September 22). MTCNN. Retrieved May 13, 2022, from https://arsfutura.com/magazine/face-recognition-with-facenet-and-mtcnn/ *FaceForensics++*. (2020, April 10). Kaggle.https://www.kaggle.com/datasets/sorokin/faceforensics

Francisca N.O. and Edward O. A,(2015) A Mitigation Technique for Internet Security Threat of Toolkits Attack, International Journal ofComputerScience and Security

(IJCSS).ww.cscjournals.org/manuscript/Journals/IJCSS/Volume9/Issue5/IJCSS-1134.pdf. Vol.9, pp225-237

Johansson, E. (2020). *Detecting Deepfakes and Forged Videos Using Deep Learning | LUP Student Papers*. Deepfake. Retrieved May 14, 2022, from https://lup.lub.lu.se/student-papers/search/publication/9019746

N. Bhakt, P. Joshi, and P. Dhyani, A Novel Framework for Real and Fake Smile Detection from Videos, 2018 Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1327-1330.

Narkhede, S. (2021, June 15). *Understanding Confusion Matrix - Towards Data Science*. Medium. Retrieved May 14, 2022, from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Saxena, S. (2021, December 11). *Convolutional Neural Network — A brief introduction*. Medium. https://becominghuman.ai/convolutional-neural-network-a-brief-introduction-c044302b3271

Team, K. (2020, April 1). *Keras documentation: EfficientNet B0 to B7*. https://keras.io/api/applications/efficientnet/

Wikipedia contributors. (2022, May 7). *Flask (web framework)*. Wikipedia. https://en.wikipedia.org/wiki/Flask_(web_framework)#:%7E:text=Flask%20is%20a%20micro%20web,party%20libraries%20provide%20common%20functions.