Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)
IEEE Xplore Part Number:CFP20OSV-ART; ISBN: 978-1-7281-5464-0

Deep Learning model-based Multimedia forgery detection

Yash Shah, Parth Shah, Mansi Patel, Chinmay Khamkar and Pratik Kanani
SVKM's Dwarkadas J. Sanghvi College of Engineering, University of Mumbai
Mumbai,India
Email:{yashshah518,parthvipulshah,mansipatel63673,khamkarchinmay4,pratikkanani123}@gmail.com,

***Abstract*— Images and videos can be spread very conveniently using social media platforms like WhatsApp and Facebook. The authenticity of this information cannot be verified easily but it spreads swiftly. Fake images or videos are a new threat for people as they spread false information and rumors. Advances in technology have given rise to several techniques that can easily generate fake images or videos. Deepfakes and spliced images are some of the results of such advances. They pose a great menace to the internet. Tackling and detecting such an entity is a tricky task. Our paper portrays a technique to detect such entities. It will assist people in detecting bogus content and have confidence on the legitimacy of the content on the internet. We present a description of CNN based approach and evaluate its results. The drawbacks of the traditional approach have been minimized using Inception Residual Network architecture based CNNs.**

***Keywords*— Deepfake, Convolutional Neural Networks, error level analysis,fake images, Inception Networks, Residual Networks**

I. INTRODUCTION

Just like the industrial revolution of the 19th century, information revolution of the 20th century the 21st century is considered to be the age of automation revolution. In this age of automation, internet is considered to be man's greatest invention. Millions and billions of data being transferred every second throughout the internet. As they that every coin has two sides, the dark side of the internet can't be ignored. Fake news and fake images are one such category of the ever-increasing list of dark arts of the internet. According to the EU High Level Expert Group (2018), false news defined as disinformation, i.e. all forms inaccurate, false or misleading information presented, promoted, or designed. Fake stuff on the internet is generally used to promote economic, political propagandas to increase the number of clicks of fake articles to get them trending which is against the interest of one party or individual. In addition to it fake news or images can also affect the price of shares which can benefit the parties who released the fake news. Various studies conducted on the spread of fake content conclude that the type of fake news which is commonly accepted are related to health, religion, financial fraud, politics, science and technology. A total of 84.5% of all respondents stated that they felt bothered by false news, and more than 70% agreed that fake news disturbs harmony community and hamper development. Apart from written form, around 40% of respondents stating that the spread of fake news also often accompanied by pictures. Historically it has been proven that a message is better conveyed through images and pictures rather than text. Humans are able to comprehend visuals better than writing. As a result, images are a powerful tool to spread fake news in today's era. Also supported by the advancements in the field of ML and AI, it has become possible to create fake images with the help of tools such as Photoshop, even non-expert users can easily modify an image and obtain realistic results. The constant improvement in the GAN's technology has also enabled to a steady rise in the number of fake images generated. In recent years the GAN's technology has also enabled common user to swap faces using specialized software like the FaceApp[18] and super impose one face on another. In technical terms this phenomenon is called as deepfakes and poses a huge threat to an individual or an organization. It can be used for malicious activities like defaming some big personality like a movie star or a president of a country. Tackling this problem is an urgent requirement in today's world. The main difference between fake images and deepfakes is that deepfakes are implemented on human faces and other facial features wheres in fake images, any object exisiting in this world can be added to an exisiting image or these objects could also be removed from the image containing in the image. Deepfakes are generated using deep neural networks such as Generative Adversarial Networks (GANs) whereas fake images are generated using PhotoShop and other similar photo editing softwares .Some of the editing techniques such as splicing [11], retouching [12], copymove [13] exist. The properties of deepfakes makes it very difficult to detect them as they appear as real images. This make it very difficult to capture the difference between deepfakes and real images using human eye. Hence special techniques to detect image forgery needs to be developed in order to determine the authenticity of an image. To achieve this purpose a lot of images with class labels as real and fake are required to train the model. With ordinary machine learning techniques, it is difficult to come up with a good model so for that reason deep learning is the right approach to tackle this problem of image forgery. In our proposed model we have used a technique that uses convolutional neural networks along with error level analysis and further tried to incorporate Residual neural networks or RNN to improve the performance of the model and achieve better results than just using normal CNN.

II. LITERATURE SURVEY

X. Yang, Y. Li et al [1] have implemented a system that exposes deepfakes using inconsistent head poses. The image is fed into a face detector. The detector extracts 68 facial landmarks using a software package. The extracted facial landmarks are compared with standard landmarks. The head poses from the centre and the whole face is estimated. The obtained differences are flattened to form a vector. The SVM classifier is fed with the resultant vector to check whether the image is fake or real.

The main principle of this method is based on the concert that when deepfakes are created by superimposing a synthesized face region on the original image and by virtue of which, it introduces errors that can be revealed when 3D head poses are approximate from the face images.

The technique is advisable when the dataset is not much large. There should not be much noise in data. This method has good accuracy.

The work proposed by A. Qayyum, J. Qadir et al [2] are based on smart contracts. This method uses multiple smart contracts to maintain the authenticity of the information. The publisher of content is verified using public/private keys (cryptography). The entire news /content is stored on the block chain. Even to add or remove the content the identity of the person is verified.

It uses semantic similarity to verify the genuineness of the content. Any content that cannot be traced to its origin is marked as fake and discarded. The actions of the agents adding contents are also verified to determine his/her genuineness. The proof of truthfulness can be determined using a Merkle tree.

The approach uses multiple smart contracts, as a result, it might become computationally heavy.

F. Matern, C. Riess et al [3] have proposed the system which has tackled 3 types of forgery image generation techniques i.e. generated faces, deep fakes, face2face. For the generated faces technique, difference in eye colour is used to detect generated faces.

The facial landmark for each image is detected along with pixels of the iris to calculate the eye colour features. In this method two checks for consistency are used to identify the failing cases in iris detection, the first one being that the centre of the iris and the centre of the eye should be somewhat similar for the both the eyes. Also the radii of both the iris must be same.

For deepfakes technique, the missing reflections and missing details in the eyes and teeth area are exploited. Again, for this method the facial landmarks are noted for the given image. The separate the teeth from the face, the image of it is converted to a grayscale image with the help of tools. The convex hull of the region of interest i.e the mouth along with its pixels are clustered with the help of K- Means clustering into bright and dark cluster. The pixels belonging to the teeth are the bright ones from the clustering performed. The work proposed by J. Kim, S. Han et al [4], detects a fake or masked facial image which has become tough to detect due to advancements made in deep learning, computer vision and image processing techniques.

The technique uses shallowNet, VGG-16 and Xception as CNN models to train the dataset. The dataset used in this approach divides the images into four parts namely normal, validation, disguised and imposture. This method separates the original image from the fake image, and they are trimmed using face coordinates. The trimmed image is then flipped horizontally and vertically via augmentation. After the images are done processing they are then classified as real or fake image using the above three models mentioned. The best accuracy was found with Xception with 62%. Its accuracy is pretty average. Number of layers in CNN models used are less. This approach is suitable with small to medium size datasets. Original and imposters images can be differentiated.

Richard Durall et al [5] have used a method which relies on classical frequency analysis of images and reveal the variation in performance of the image at higher frequencies. Techniques such as Generative adversarial (GAN) and Variational Autoencoders (VAE) are predominantly used to generate fake images. Accuracy for high, medium and low-resolution images are 100%, 96% and 90% respectively.

This technique is advised when there’s need to have better accuracy when less amount of dataset available.

A fake feature network-based pairwise learning method has been proposed by Chih-Chung Hsu et al [6]. A deep learning-based approach called as contrastive loss is used to detect imposter images from the real ones.

An input of pairwise information is passed into a two streamed network which is developed using the DenseNet architecture. The detect the features between real and fake image, a common fake feature network is trained using the pairwise learning approach. Finally, to detect if the image is fake or real, a classification layer is attached to the given networks of fake feature.

The results obtained from this approach drastically improves the application of detection compared to other image detectors.

This approach must be used when the dataset is relatively small with major manipulations to the original image.

First one is to convert the RGB coloured image to YCrCb coloured image. Next is to apply discrete cosine transform (DCT) to Y component. Then apply the quantization according to JPEG quantized table. Finally apply Huffman encoding on it.

The CNN model used has a sixty-four-layer filter of size 3 x 3 with ReLU as the activation function and a max pooling layer. The size of the pool is 2 x 2 is used.

The main problem with previous methods in this domain was the area of detection in such compressed areas are not properly aligned. This infers that we couldn’t calculate features in a sliding window to detect compressed area. This problem is tackled in this paper. The sliding window step of 5 is considered because this kind of step can get a JPEG alignment grid. The training and testing dataset is divided in a ratio of 70:30. The test accuracy of 0.95 during 300 epochs of the CNN training is reached.

This approach can be used with a large dataset and high accuracy is expected.

Pakpoom Mookdarsanit et al [8] proposed a method to detect image forgery by comparing the truth values of XOR between two images and determinant of 3x3 pixels increasing the speed by 36.42% than other techniques. Algorithm goes as follows:

1. Find the pixel and its neighbours.

2. Compute the Euclidean of RGB vectors3. Compute determinant of pixels 4. Compare determinant of pixels between two images. This method can also be integrated with trigonometry for solving the geometric correction such as rotated images. This technique can be used where there’s major manipulation in images.

The implementation proposed by P. He, H. Li et al [9] is primarily based on color models (RGB, YCbCr, HSV, Lab) used in various multimedia. The approach is based on a fact that there appear inconsistencies in images that are generated using GANs and the images that are original. There is some mismatch in chrominance of an image that has been captured by a camera and an image that is fake. Each image is converted into a residual signal form of its chrominance components. CNN is fed with resultant residues to get the better representations. A random forest algorithm is used where the obtained representations are fed into it to determine the credibility of the images. The proposed model performs better than other modern models. Also this technique has a more robust approach to detect accuracies against post processing attacks especially for blurred images.

This technique can be used for an accurate transmission if the complexity is acceptable. The accuracy achieved with this technique was 90% for fake image, 60% for deep fake and the combined accuracy was 55%. This technique relied on lossy compression and required image to be compressed. Also, extra step of preprocessing was required and some approaches were dataset dependent. In order to eliminate the drawbacks a residual network-based approach is presented in the paper. The proposed approaches is based on a combination of Inception and residual networks. The literature review is summarized in the table 1.

Table 1 Summary of literature review

Paper reffered	Approach used	Pros	Cons	Comments
Exposing Deep Fakes Using Inconsistent Head Poses	SVM classifiers.	Provides good accuracy. Flexible approach	May not be suitable for large datasets Requires clean data.	This technique is advisable when the dataset is not much large. There should not be much noise in data. This method has good accuracy.
Using Blockchain to Rein in the New Post-Truth World and Check the Spread of Fake News	Blockchain based Smart Contracts	It is a robust approach. The content once added cannot be modified.	The approach uses multiple smart contracts, as a result, it might become computationally heavy.	This approach is suitable for preventing the spread of deepfakes. It can be combined with other AI/ML techniques to increase its usefulness.
Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations	K Nearest Neighbor Classifier	Easy to implement on small datasets. Fast approach.	Difficult to implement on large datasets. Accuracy is less than deep learning models.	This technique can be used when the dataset is not large and time is a constraint to achieve decent results.
Classifying Genuine Face images from Disguised Face Images	Convolutional Neural Network (CNN)	Easy to implement on small datasets. CNN models used for training the datasets.	Difficult to implement on large datasets. Accuracy is pretty average.	This approach is suitable with small to medium size datasets. Original and imposters images can be differentiated.
Unmasking DeepFakes with Simple Features	Classical Frequency Domain Analysis	No need for a large dataset. Accuracy is high.	Less accuracy on low-resolution images. Loss of temporal information.	This technique is advised when there's need to have better accuracy when less amount of dataset available.
Deep Fake Image Detection Based on Pairwise Learning	Pairwise Learning	Best suitable for minor as well as major manipulations in images.	Large amount of annotated data is required. Performs well on benchmarked datasets, but can fail badly on real world images outside the dataset.	This approach must be used when the dataset is relatively small with major manipulations to the original image.

III. Research Gap

In the case of deep fake detection, initially we achieved an accuracy of 52%, but on increasing the epochs from 10 to 30 we achieved the accuracy of 81%. When general image classification is considered, the current system gives accuracy of 93%. The CNN based architecture performs poorly when trained on a dataset of deepfakes. The performance degrades further when the dataset contains a combination of deepfakes and fake images. A robust model with an accuracy of over 90% was possible by using a residual network based architecture. The residual nature of the CNN prevented feature loss over the course of training and resulted in superior accuracy. The model also improved accuracy over the previous approach when trained on a combination of deepfakes and fake images.

IV. Objective

The main objective of our project is to develop a deep learning model to detect fake images of various types such as image manipulation, image overlay, image forgery, deep fakes etc. all over the internet. Due to advancements in the field of machine learning, various tools are available to common people to create fake images and spread it across the internet and the rate at which fake images and news being spread in growing exponentially and difficult to stop.

It can lead to a serious level of defamation as people tend to make conclusions without checking the facts or origin of the source. To control the rise of fabricated stuff published on the internet, our objective is to develop an agent which can detect the credibility of the image and tell if its fake or real to the user. The objective of the project is to have a user-friendly interface such as a user will upload an image and the output of it will be either fake or real based on our analysis.

V. MOTIVATION

Earlier the world faced a big problem due to fake news, but now there is even a bigger problem "Deep fake" and Forged images. It is a kind of exploitation of AI and Machine Learning technology used to perform face swap thus creating a mirage that someone said something which in real they didn't say or are someone they're not and such changes are not perceptible to human eye compelling people to believe easily.

Deep fake exponentially affects celebrities, folks and politicians; it leads to defamation, Intentional infliction of distress, Breach of confidence and public disclosure of private facts, false light, Impact on business and also affects privacy.

Fig. 1 Samples of Deep fake testing dataset [3]

The stories depend on images to sell bogus records. The people publishing and promoting fake news routinely take photos out of context, digitally alter them, or combine them with text to influence readers, knowing that people lean towards accepting photographs as truthful representations.

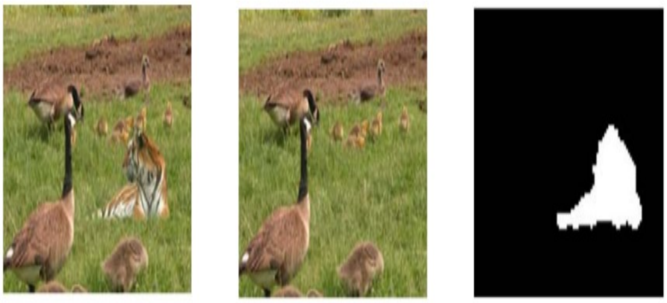

Fig. 2 Left: Original image, Middle: Tampered image, Right: Detected Region [10]

CREO point's Golden Stein explains that "That's a serious problem since AI can't reliably detect fake news or fact check fast enough". All in all, the problem is rapidly increasing and has an inadequate impact on the world. Considering such a threat these images and videos pose, we try to reduce it by our research work.

VI. PROBLEM DEFINITION

A lot of images are shared on social networking sites such as Facebook, Instagram, and WhatsApp. These images become the evidence based on which people may infer various things. Due to the increase in use of image processing software, it becomes very easy to manipulate images. This is very serious concern for all the social networking sites as these forged images can become source of rumors and can also result in some mishaps. It is very important to verify the authenticity of these images to prevent the intent behind its spread. Considering the seriousness of this ongoing issue, we came up with the solution to detect forgery in the images. For this, we worked on detecting manipulations in general images and focused to detect forged faces. In the case of deep fakes, we have used Keras Image Data Generator class which is used for classification. Over the last few years, Deep Learning has shown excellent results when it comes to Computer Vision. This is the reason why we have used Deep Learning in the form of CNN to classify images as original and fake and ELA (Error Level Analysis) to detect the manipulation in images. Error Level Analysis is the technique which detects forgery based on ratios of compression level of images. CNN is the type of network where the information flow is unidirectional i.e. from input to output. For image classification, we gave images as input to the CNN network so that every pixel can be processed. ELA is dependent on lossy compression of images and may not function well for images that are not compressed. The technique does not produce convincing results for deepfakes and hence it cannot be used to detect deepfakes. The convolutional layers can be made more effective by incorporating a residual network architecture. The residual architecture feeds inputs from the layer two stages before the previous stage, this prevents feature loss and ensures that the network has a knowledge of features extracted in the previous layers. Along with the Residual network-based architecture an Inception network allows the network to try out different combinations of filters and pooling layers to determine the best possible combination of feature extractors. This allows the network to train on the best combination of filters. The model for forgery detection is based on a combination of inception and residual networks known as Inception Resnet. This combination allows the network to take advantage of useful characteristics of both the approaches.

VII. METHODOLOGY

A. Error Level Analysis

Error level analysis is a forensic technique to detect fake images. It is based on lossy compression of images. A loss compression of an image is an irreversible compression method to encode image and reduce its size. An image that undergoes lossy compression cannot be decompressed to get the original image back. When an image undergoes a lossy compression such as JPEG the compression level at each pixel is the same. The JPEG algorithm works as follows:

- Divide the image into grids of size 8*8 pixels
- Compress the image on a scale of 10:1

For the images not altered digitally the compression level for all the grids should be same. The level of degradation for each square should be the same. For the images that are digitally modified the compression rate will not be the same for all the grids. There will be a mismatch in the compression level of the grids.

Error level analysis resaves the image at 95% compression level. Now, it evaluates the difference between the original and the resaved image. If the image is not modified the difference will be the same across all the grids. However, if the image is modified the difference will not be consistent across all the grids. Thus, this difference in compression shows suspected areas of forgery in the image.

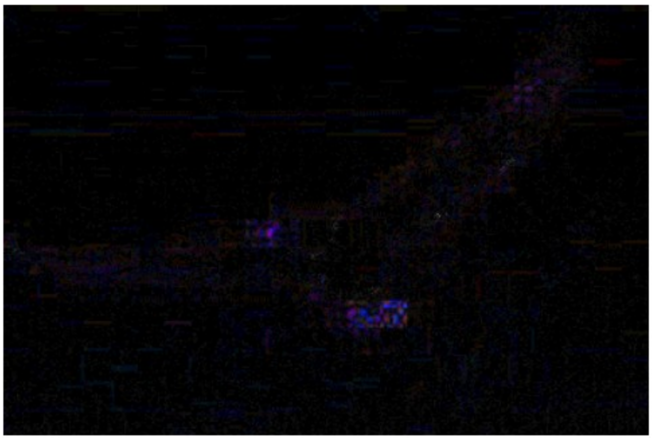

Fig. 3 Result of ELA on a real image

Figure 3 shows the result of Error level analysis on a real image. The image has not been tampered or photoshopped. The image was a picture of a bird. As seen in figure 3 performing ELA on the image produces almost black image and colored regions representing the head and tail of the bird. This uniformity in Figure 3 conveys that the image is genuine and is not forged.

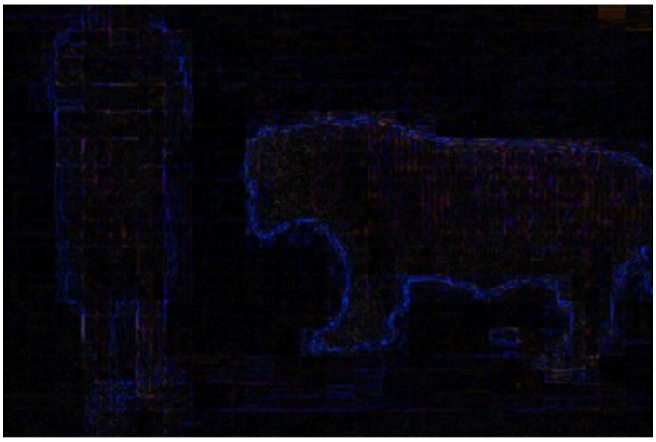

Fig. 4 Result of ELA on a tampered image

Figure 4 shows the result of ELA on a forged image. The difference between figure 3 and figure 4 is clearly visible. In figure 4 there is not much uniformity in the image. The blue shaded shape reveals that there is some difference in the compression levels at that place in the image. This reveals that the shape of an animal indicated by the blue shade belongs to some other image. This reveals that the original image did not contain the animal. The animal was forged into the image. Thus, ELA helps to identify the changes in compression to detect forgery.

B. *Convolutional Neural Networks(CNNs)*

A Convolutional Neural Network is a deep learning algorithm suitable for computer vision. It can take images as input and assign importance to various aspects of an image on its own. As a result, it can differentiate between various images on its own. CNNs are useful in finding high level features that may not be visible to the human eye. The architecture of the CNN is analogous to the visual cortex of the human brain. A CNN performs better than a multi-layer perceptron as it can handle complex images well at a decent accuracy. A CNN can easily capture spatial and temporal features in an image by application of pertinent filters. As a result, it can fit the characteristics of an image into a model. The role of a CNN is to reduce an image in a form such that it becomes simpler to process it.

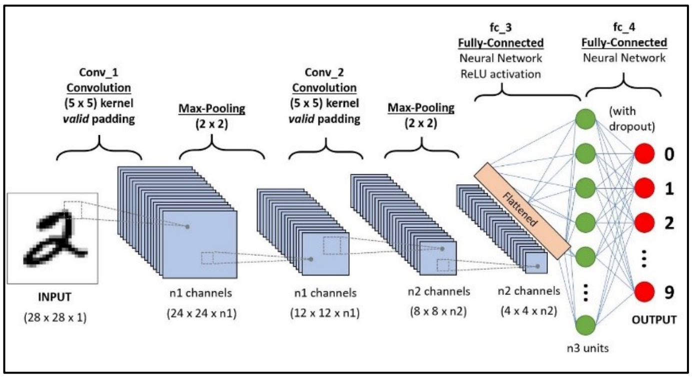

Fig. 5 Architecture of CNN [20]

The architecture of a CNN is as follows:

- Input layer: The image to be processed is fed into the neural network at this layer
- Convolutional layer: It takes the image and creates and tries to identify the feature of the image. It now understands the features and creates a list of features that can be used for further processing the image
- Pooling layer: It scales down the information generated in the convolutional layer, keeping the most essential features required. It reduces the spatial size of the convolutional layer so that processing the image becomes computationally lighter.
- Fully connected input layer: Flattens the image into a single column vector
- Fully connected layer: Applies weights to the feature vector generated by the previous layer
- Fully connected output layer: It calculates probabilities for class labels so that the image can be classified.

C. *System Architecture for fake image detection*

The system design for classifying the image as fake or real is shown below. It employs CNN and error level Analysis to detect fake images as well as deep fakes.

- Dataset: Kaggle Casia Dataset [24]
- Data pre-processing

- The images are subjected to ELA and stored in a directory.
- The processed images are resized to a grid of dimension 128 x 128 x3 where the height represent three different chrominance components viz. R, G and B
- The resized image set is split into testing and training data with a ratio of 80% to 20%

- CNN model:
  - First convolutional layer with 32 filters to extract features
  - Second Convolutional layer with 32 filters to extract features from the first convolutional layer
  - A max pool of dimensions 2x2x32 to flatten the features to a vector.
  - The activation function used at the output is SoftMax algorithm to classify the image as fake or real
- Samples used:
  - Training: 3768
  - Testing: 943

D. *Result analysis for fake image detection system using ELA and CNN*

The model resulted in an accuracy of 91% on the testing data. The confusion matrix for the model is shown below:

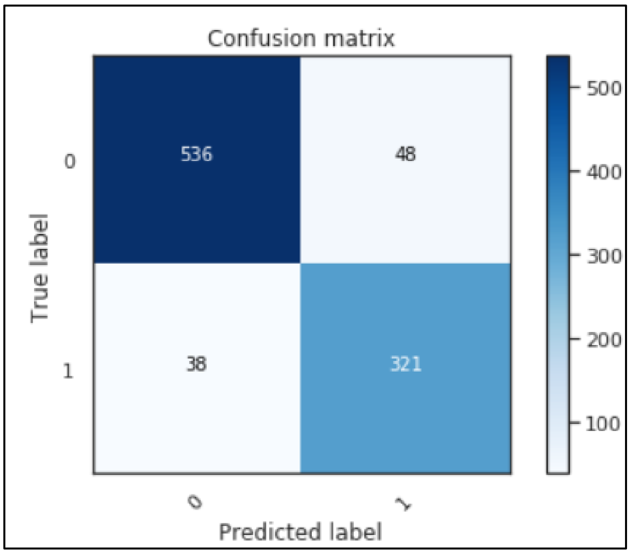

Fig. 6 Confusion Matrix for model

The ELA technique is only suitable for images that are subjected to lossy compression techniques. If an image that has not been compressed or has been subjected to lossless compression, then the ELA technique may not work properly. When dataset contained deepfakes, the model did not perform well and resulted in an accuracy of 60 %. When the dataset contained a combination of deep fakes and fake images the performance further degraded and resulted in an accuracy of 55%. Hence, the model is not capable for detection of deepfakes.

E. *Inception Residual Networks*

Inception Residual networks are a novel advanced type of CNNs. The network is a combination of Inception and residual networks.

Residual network is a special type of neural network. It is a newer version of CNNs. In this type of network, the input state of a previous layer is added to the current active layer. This addition of the previous state helps to preserve the features learnt by the previous layer and prevents feature loss. The earlier approach of using pure CNNs could not preserve features while training and resulted in feature loss. This resulted in performance degradation for deep fakes. The residual network can preserve features and is capable to be employed for deepfakes.

Figure 7 shows a detailed diagram of a residual network. As shown in fig. 7 the input is fed to the Convolution layer where convolution operation is performed. After pooling the convoluted characteristics, the result is subjected to ReLU activation. The same sequence of operations is performed again, and the result at this stage is added along with the input to the first convolution layer in the block. This result is passed on to the successor block after ReLU activation. This preserves the features from previous layer and results in greater accuracy.

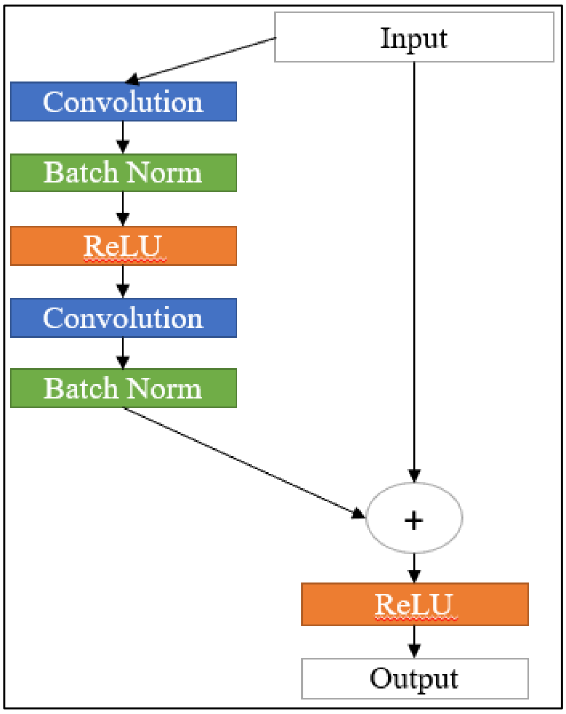

Fig. 7 Basic Block of residual network

The inception network is kind of network that allows the model to choose the best possible dimensions for the convolution filter or to choose a pooling layer either. For an iteration through the inception network produces output for each type of convolution layer and / or pooling layer that is desired to be tested to determine the best possible combination for the model. The outputs for each operation are concatenated and stacked together. It is the network that decides which combination is the best fit for the model. It will try out different combinations and choose the best one. The basic block of an Inception Network is shown in figure 8. One may wonder that the amount of computational power required will significantly increase in order to carry out different combinations. The computational cost can be significantly reduced by introducing a bottleneck layer before the combination of layers that are to be tested. The bottleneck layer consists of a convolutional filter of dimension 1 X 1. This reduces the computational cost to 1/10th of the cost required to carry out all the combinations. The reduction in size by the bottleneck layer does not affect the performance of the layer if positioned properly in the network.

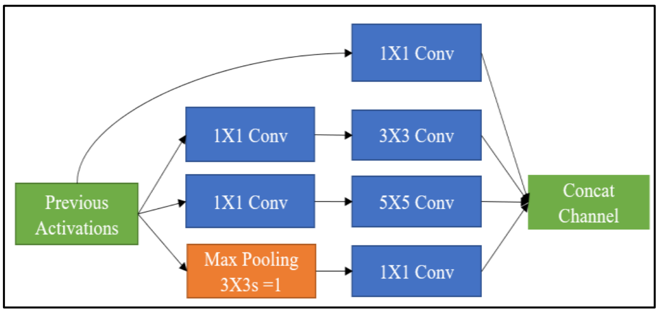

Fig. 8 Block of Inception Network

An inception residual network is a combination of both the two approaches. The architecture of the model to classify images as fake or real is based on Inception Resnet V2. Inception Resnet V2 is defined by Google under the category of Google Net models. The model has been trained previously on a million of images from the ImageNet dataset. The dataset is capable to classify an image into a variety of object types like keyboard, books, and a plethora of animals. The model is 164 layers deep. The network consequently has learnt a variety of useful features for classification. The model is an improvement over the Inception V3 architecture proposed by Google. The model has a superior accuracy and is a backbone of many computer vision applications. The inception network can hence be trained on a dataset of deepfakes and spliced images to classify them as genuine or fake.

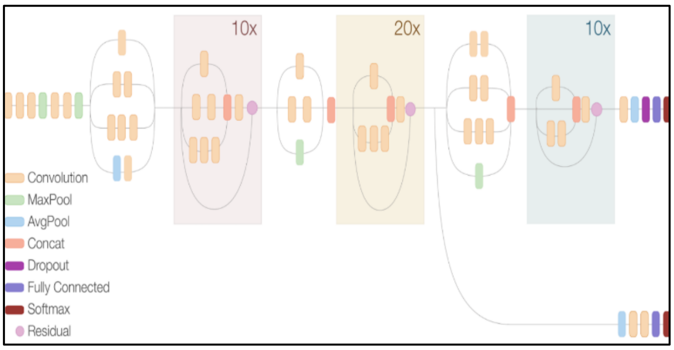

Fig. 9 Inception Resnet V2 Architecture [22]

Figure 9 describes the architecture of the Inception Resnet V2 model. The model is a combination of different elements like convolutions, pooling, Residual blocks, SoftMax Activation and fully connected layers. Each layer in the model is displayed using a particular colour. The legend in the figure 9 shows the colour for each category of block in the model. The way each block is connected to the other is shown in the figure 9. The network starts with a combination of Convolution operations followed by Max Pooling layer. An inception block containing a different combination of Convolution and pooling blocks, follows the max pooling block. The succeeding sequence of blocks is repeated 10 times and is connected to an Inception Block. The next sequence of blocks is repeated 20 times and is again followed by an Inception Block. The next sequence of blocks is repeated 10 times and is continued by a convolution block followed by a fully connected layer and this in turn is connected to a SoftMax activation to produce the final output.

F. Proposed Inception Resnet V2 based model

Inception Residual Network V2 model can be used for detecting deepfakes. The model has already learnt a bunch of useful features as a result of pre-training. This makes it very useful for computer vision applications. The architecture of the model proposed is same as figure 9. The various parameters for the systemare:

- Dataset: Kaggle Deepfake detection challenge and Fake face Detecion [25]
- Data pre-processing: Not much preprocessing required just reshaping the images to 128 X 128 X 3
- Samples used:
  - Training: 2996
  - Testing: 749

G. Result analysis of the proposed methodolog

The Inception Resnet V2 model resulted in an accuracy of 92% when trained on a dataset of deepfakes. The confusion matrix is shown below:

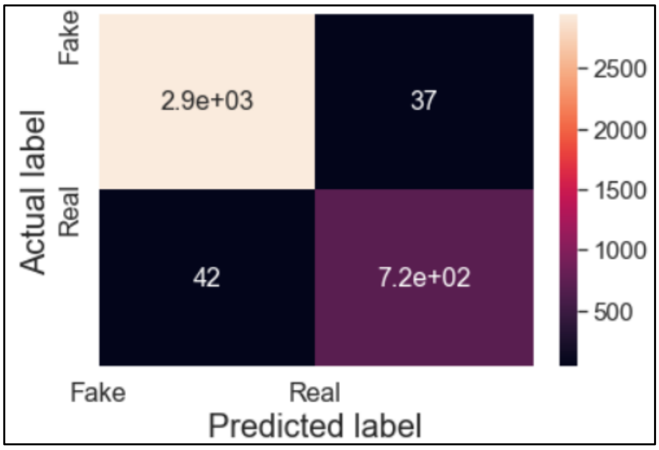

Fig 10 Confusion matrix for proposed methodology

The Confusion matrix shows the efficacy of the proposed methodology. The true positive count is 2949. The number of false positives is 37. The number of true Negatives is 720 and the number of false negatives is 42. The various evaluation parameters are:

- Accuracy: 92%
- Precision: 94%
- Recall: 95%

The dataset performs equally when trained on a dataset of Fake images. The model resulted in an accuracy of 91%. The model also results in a better performance when trained on a combination of deepfakes and fake images and the resultant accuracy is 79% which is a considerable improvement as compared to the previous approach. The model also eliminates the need for rigorous preprocessing as compared to the previous approach where each image had to undergo ELA. The only pre-processing required is to reshape the image to the desired format. The Residual nature of the proposed method prevents feature loss during training which is an improvement over the previous CNN based approach. The inception nature of the method helps to determine the best combination of filters to use for better accuracy. The earlier approach was static and the model was not capable of detecting choosing the best filter for training. Table 2 shown below summarizes the performance of the two approaches.

Table 2 Comparison of the proposed and previous approach

Approach	Fake images (accuracy)	Deep Fakes (accuracy)	Combination of fake images and deepfakes (accuracy)
CNN and ELA	90%	60%	55%
Inception Resnet V2 (proposed)	91%	92%	79%

VIII. CONCLUSION

We have successfully implemented the image forgery detection system with accuracy of 91% for deep fakes using the Inception Resnet V2 architecture. The ELA technique performs well on fake images but not for deepfakes or a combination of deepfakes and fake images. The Inception Network performs exceptionally for deep fakes and performs convincingly for a combination of deepfakes and fake images. The approach is not dataset dependent and also performed well when tested on images from a different dataset.

IX. FUTURE SCOPE

The current model works well with images but is not compatible with videos. Also, efforts are being taken to highlight suspected areas of an image. The ways to improve the accuracy of the combined model will have to be evaluated. By making changes to the model or by performing some preprocessing before training the model can help further improve the accuracy. The model performance can be optimized by customizing the network architecture from the GoogleNet version. Efforts will be made to extend this model for functioning on deepfake videos with a good accuracy. Videos can be detected by breaking them into frames and then individually detecting if the frame is real or fake. This approach will split the video into an array of images and the images can be fed to the classifier to detect forgery. This approach will make the scope of the model broader and hence work on a variety of images and videos.

REFERENCES

[1] X. Yang, Y. Li and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 8261-8265.

[2] A. Qayyum, J. Qadir, M. U. Janjua and F. Sher, "Using Blockchain to Rein in the New Post-Truth World and Check the Spread of Fake News," in IT Professional, vol. 21, no. 4, pp. 16-24, 1 July-Aug. 2019.

[3] F. Matern, C. Riess and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 2019, pp. 83-92.

[4] J. Kim, S. Han and S. S.Woo, "Classifying Genuine Face images from Disguised Face Images," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 6248-6250.

[5] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, Janis Keuper, “Unmasking DeepFakes with Simple Features,”2020 arXiv:1911.00686v3 [cs.LG], Fraunhofer ITWM, Germany.

[6] Chih-Chung Hsu, Yi-Xiu Zhuang, Chia-Yen Lee, “Deep Fake Image Detection Based on Pairwise Learning,”2019 Department of Management Information System, National Pingtung University of Science and Technology, 1, Shuefu Road, Neipu, Pingtung 91201, Taiwan.

[7] A. Kuznetsov, "A New Approach to JPEG Tampering Detection Using Convolutional Neural Networks," , Novosibirsk, Russia, 2019, pp. 0520-0524.

[8] Pakpoom Mookdarsanit, Lawankorn Soimart, Mahasak Ketcham, Narit Hnoohom, “Detecting Image Forgery using XOR and Determinant of Pixels for Image Forensics”, 2015, Chandrakasem Rajabhat University, Bangkok, Thailand.

[9] P. He, H. Li and H. Wang, "Detection of Fake Images Via The Ensemble of Deep Representations from Multi Color Spaces," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 2299-2303.

[10] Weiqi Luo, Jiwu Huang and Guoping Qiu, "Robust Detection of Region-Duplication Forgery in Digital Image," 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 2006, pp. 746-749.

[11] H. Farid, “Exposing digital forgeries from JPEG ghosts,” IEEE Trans. Inf. Forensics Security, vol. 4, no. 1, pp. 154–160, Mar. 2009.

[12] J. Li, X. Li, B. Yang, and X. Sun, “Segmentation-based image copymove forgery detection scheme,” IEEE Trans. Inf. Forensics Security, vol. 10, no. 3, pp. 507–518, Mar. 2015.

[13] G. Cao, Y. Zhao, R. Ni, and X. Li, “Contrast enhancement-based forensics in digital images,” IEEE Trans. Inf. Forensics Security, vol. 9, no. 3, pp. 515–525, Mar. 2014.

[14] M. Đorđević, M. Milivojević and A. Gavrovska, "DeepFake Video Analysis using SIFT Features," 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia, 2019, pp. 1-4, doi: 10.1109/TELFOR48224.2019.8971206.

[15] Vijayakumar, T., and R. Vinothkanna. "Mellowness Detection of Dragon Fruit Using Deep Learning Strategy." Journal of Innovative Image Processing (JIIP) 2, no. 01 (2020): 35-43.

[16] Manoharan, Samuel. "Image Detection Classification and Recognition for Leak Detection in Automobiles." Journal of Innovative Image Processing (JIIP) 1, no. 02 (2019): 61-70.

[17] A. Khodabakhsh, R. Ramachandra and C. Busch, "Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content," 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, 2019, pp. 1-6, doi: 10.1109/QoMEX.2019.8743316.

[18] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[19] Y. Li, M. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630787.

[20] A Comprehensive guide to convolutional Neural Networks [Online], Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[21] Sam Gossamer and Michael Wiber “Training and inestigating residual nets” http://torch.ch/blog/2016/02/04/resnets.html

[22] Alex Alemi “Improving Inception and image classification in TensorFlow” https://ai.googleblog.com/2016/08/improving-inception-and-image.htm

[23] Wireless Lab. (2016, December). FaceApp – AI Face editor. Retrieved September 02, 2020 from https://faceapp.com/app.

[24] Casia dataset, Kaggle [Online] https://www.kaggle.com/sophatvathana/casia-dataset

[25] Deepfake detection challenge, Kaggle [Online] https://www.kaggle.com/c/deepfake-detection-challenge/data