Deepfake Image Detection Using Convolutional Neural Networks

[1]Bondalakunta Praditha,[2]Rohan Sanjay Mahajan,[3]Vaishnavi VR,[4]Vibha T Karunakara,

[5]Prof. Deepa Pattan

[1]Student,[2]Student,[3]Student,[4]Student,[5]Assistant Professor
[1]Department of Information Science and Engineering,
[1]Sai Vidya Institute of Technology, Bengaluru, India

Abstract - Deepfake images, which are images that are manipulated or synthesized using deep learning techniques, have become a serious concern for society as they can be used to spread disinformation or to deceive people. With the rapid advancements of deep learning and computer vision technologies, the creation and dissemination of deepfake images have become a serious for society. In this paper, we propose a deep learning-based solution using Convolutional Neural Networks (CNNs) for detecting deepfake images. Our proposed method extracts features from images using a CNN-based model and uses them to classify images as either authentic or manipulated. A large dataset of deepfake images was used to evaluate the approach and show that it achieves high accuracy and robustness to various types of image manipulations. Our proposed method can be applied to various real-world scenarios, such a social media platform, to prevent the spread of deepfake images and to ensure the authenticity of visual content.

Index Terms – Convolutional Neural Networks, Deepfake Images, Deep Learning, Tensorflow, Image Classification

I. INTRODUCTION: DEEP FAKE AND THE NEED OF DETECTION

The rise of deepfake technology has brought both excitement and concern to various industries including media, entertainment, politics. Deepfakes are generated using advanced machine learning techniques, particularly Generative Adversarial Networks (GAN) to create realistic but fraudulent image content that can deceive human perception. These deep fake pose significant challenges in terms of their potential to spread misinformation, create fake identities and manipulate public opinion. Therefore, the need for robust and reliable deepfake detection techniques is of paramount importance to safeguard against the malicious use of this technology.

(1) *Deep fakes*

In recent years, CNNs have emerged as a powerful tool for deepfake detection due to their ability to automatically learn features from large datasets, which enables them to identify patterns in visual content. CNNs are a type of artificial neural network that can process data in the form of multiple input channels, such as images and learn to extract relevant features from the input data through Convolutional, pooling and fully connected layers. These learned features can classify the input data as genuine or fake. The rapid increase of deepfakes poses significant concerns regarding the authenticity of news disseminated by the mass media, as well as potential threats to politics, companies. To address this alarming scenario, the development of effective tools for deepfake detection has become necessary [11]. Recognizing the gravity of this situation, prominent companies such as Meta, Pinterest and Microsoft have taken proactive steps to combat deepfakes. These companies have established a database of fake videos to facilitate research on novel detection techniques, while these companies have jointly joined the deepfake detection initiative. These efforts demonstrate the growing recognition of the need to unmask or detect deepfakes and mitigate their potential adverse impacts [12].

(2) *The need for detection*

- Misinformation: Deepfake technologies are used to create false or misleading information, which can be spread online, potentially damaging reputations or cause social or political unrest.
- Fraud and Cybercrime: Deepfakes are used to commit various forms of fraud, such as impersonating individuals or creating fake financial transactions. [13]
- National Security: Deepfake technologies create convincing fake images of government officials, potentially causing political or social instability.
- Online safety: Deepfakes create obscene or offensive material, which are used to harass or blackmail individuals.
- Legal and ethical concerns: The use of deepfake technology raises numerous legal and ethical concerns, such as issue of consent, privacy, and intellectual property.
- Protection of authentic media: Deepfakes can undermine trust, hence there is a need to protect the authenticity of genuine images.
- Preventing misuse of technology: Deepfake detection can help to prevent misuse of this technology for malicious purposes.
- Existing unreliable detection methods: As deepfake technology becomes more sophisticated, there is a need for reliable and accurate detection methods to differentiate between fraudulent and authentic media.

II. LITERATURE REVIEW: RELATED WORK ON DEEPFAKE DETECTION

Xinyi Ding, Zorhreh Razieiy. et al [1] aimed to develop a technique for detecting swapped faces using deep learning. To do this, they utilized a custom data set that they created which is now publicly available. The deep learning model developed by the researchers can provide predictions and accuracy rates for each prediction. However, the model was found to have lower accuracy rates when compared to human subjects

Scott McCloskey and Michael Albright [2] proposed a technique for detecting GAN-generated images using saturation cues. The main method they used was an SVM classifier, they trained and tested the model the dataset called Image-Net. Their model achieved an accuracy rate of 92%, which is promising for detecting GAN-generated images. This finding highlights the need for continued research and development in the area, as GAN-generated images are becoming increasingly prevalent and sophisticated.

Digvijay Yadav and Sakina Salmani,[3] presented a survey of facial morphing or forging techniques by using Generative Adversarial Networks (GANs). The authors reviewed two techniques which are Convolutional Networks along with Long Short-Term Memory. The authors also used two datasets in their study: Face2Face and Reddit user deepfake dataset. Their evaluation showed the accuracy technique using GAN-based techniques for facial forgery was 92%.

Mingzhu Luo and Yewei Xiao [4] proposed a multiscale face detection model based on Convolutional Neural Network (CNN). They evaluated the model on three datasets – CelebA, AFW and FDDB – and analysed the results. They achieved a good accuracy for discrete data. However, for continuous data the accuracy only 74%. The authors concluded that the model was effective in detecting faces at multiple scales and in handling complex image backgrounds.

C. Hsu, Y. Zhuang and C. Lee [5] proposed a method for detecting deep fake images using pairwise learning. They experimented with various techniques such as CFFN, RNN and GAN with the dataset they used being the celebA dataset consisting of over 200k celebrity images. GAN technique, which is typically used for image generation was also employed for detection. Their model exhibited superior performance in terms of decision and recall rate compared to others methods.

Xinsheng Xuan, Bo Peng. et al [6] conducted a study on the generalization of generalized image forensics. They used convolutional neural networks to detect images generated by GAN and evaluated their method on the celebA dataset, which contains over 200k celebrity images. Their proposed method was more effective than existing one, but they only achieved preliminary results. Their study provides insights into the use of CNN for detecting GAN-generated images, further research is needed to improve the accuracy and generalize the results to other datasets.

Ranjan, Sarvesh Patel and Faruk Kazi [7] conducted a study on improving the generalizability of deepfake detection using transfer learning-based CNN framework. They utilized LSTM and CNN techniques and evaluated their approach on three popular datasets-Face Forensics++, Celeb-DF, Deepfake detection challenge. Through their research they attained an accuracy of 86.49% using transfer learning and accuracy of 79.62% without transfer learning. This work represents an improvement in deepfake detection and demonstrates the efficacy of transfer learning in this area.

Shivangi Aneja and Matthias Niesser [8] proposed a survey paper that uses a dataset called FaceForensics++ which is a widely used datasets for deepfake detection research. The proposed model achieved state-of-the-art performance in both zero-shot and few-shot scenarios. The results indicate that the proposed framework can generalize well to unseen facial forgery manipulations with few examples which is a crucial feature for practical usage.

Siwei Lyu et al [9] explores the current use of deepfake detection and the challenges that need to be addressed. The author discusses the limitations of the current detection methods including their reliance on specific type of deepfakes and the need for large amounts of training data. The paper also suggests potential solution to these challenges, such as developing more generalisable detection methods and creating more diverse datasets. Overall, the paper highlights the need for continuous research and development of deepfake detection field to combat the growing threat of malicious use.

H.S Shad [10] conducted an analysis to compare deepfake detection methods with Convolutional Neural Network. They used the Kaggle dataset, which includes 140000 true and false images of people. The paper explores different CNN architectures and even creates a custom CNN model. The accuracy achieved was high. However, the recall rate was not up to par overall the paper provides a comprehensive analysis of deepfake detection using CNN.

III. PROPOSED METHOD: CNN-BASED APPROACH FOR DEEP FAKE IMAGE DETECTION

The proposed method for detecting deep fake images involves using a Convolutional Neural Network based approach. CNN is most used in tasks that require image classification, and their outstanding performance in diverse computer vision applications is well documented. [14] The proposed CNN-based approach consists of multiple layers which are of the following types, Convolutional, pooling, fully-connected layers for classification as shown in the figure 1. The figure 1 depicts the general architecture of CNN with two convolutional and pooling layers followed by one fully connected layer and the output layer.

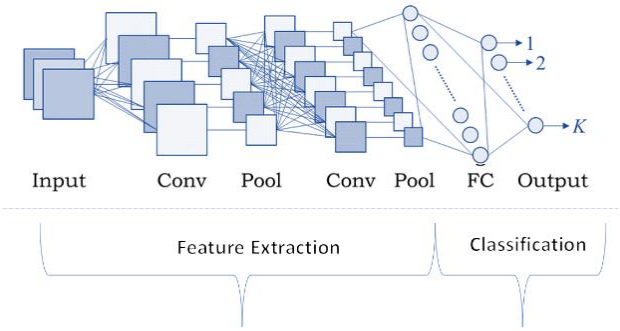

Fig.1 CNN Architecture

The first step involved extraction of key details from the received image. The layers that is followed by the input layer in the CNN architecture perform this task by applying a set of filters to the input image. The filters used in the CNN architecture capture various characteristics of the image. The output from the convolutional is passed to the pooling which maps the characteristics and reduces number of parameters in the network. The next step is to classify the input image as either real or fake. The fully connected layers in the CNN architecture perform this task by mapping the extracted features to the corresponding class labels.[19] The result from the fully connected layer is given to a SoftMax function which normalized the output to obtain the class probabilities. To train the CNN-based deep fake detection model, a big dataset of true and false images is used. The model is trained using backpropagation and stochastic gradient descent. During training, the model adjusts the parameters of the CNN architecture to minimize the cross-entropy loss between the predicted and true class labels.

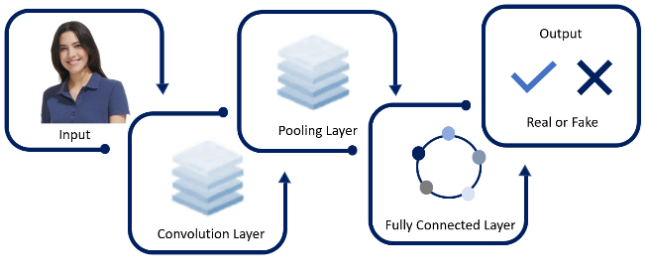

Fig. 2 Workflow Diagram

To analyze the working of the model that we built various methods have been used such as F1 score and accuracy. The performance of the model is analyzed in comparison with existing deepfake detection methods. The results demonstrates that our custom CNN-based approach outperforms existing methods in respect to various measures. As depicted in the figure 2, the proposed CNN-based approach for detecting deep fake images is a promising technique that can effectively distinguish between real and fake images and classify them using fully connected layers. The input data passes through the convolutional and pooling layers to extract and analyze features, after that the fully connected layer draws conclusions from all the data and then produces an output in the form of a binary classification to detect is the image is fake or real.

IV. DATASET AND EXPERIMENTAL SETUP: DESCRIPTION OF DATASET AND EVALUATION METRICS

In this chapter, we provide a detailed description of the dataset and the experimental setup used in our study on deep fake detection using Convolutional Neural Networks (CNNs).

(1) *Dataset*

We utilize a publicly available dataset from Kaggle, which comprised a diverse mix of 1,40,000 images. The dataset included both real and fake images, which were sourced from various sources. The real images were obtained from the public image databases, while the fake images were generated using a variety of deep fake techniques, including face-swapping, face-morphing and facial expression synthesis. The dataset was curated to ensure a balanced representation of both real and fake images, with each category having equal

representation. The size of the dataset is of around roughly 2 Gigabytes and it easy very simple to obtain it. Creating a simple Kaggle account and specifying the purpose of download allows you to access the entire database. The figure 3 displays a randomly pulled set of images from the training set. Here we can observe that each image has a label to which depicts if the image is fake or real which will help us to classify images from our testing set and other uploads.

Fig. 3: Representation of dataset

(2) *Data Split*

To ensure that the CNN model was robust and reliable, the dataset was divided into three sets: training, testing and validation. The training set consisted of 100k images, with an equal distribution of 50k true and 50k false images. The testing set had 20k images, with an equal distribution of 10k true and 10k false images. Finally, validation set also contained 20k images, with an equal distribution of 10k true and 10k false images.

```
Total training images REAL: 50000
Total training images FAKE: 50000

Total test images REAL: 10000
Total test images FAKE: 10000

Total test images REAL: 10000
Total test images FAKE: 10000


Found 100000 images belonging to 2 classes.
Found 20000 images belonging to 2 classes.
Found 20000 images belonging to 2 classes.
```

Fig. 4 Data split representation

As seen in the figure 4 multiple directories was used to store data of a certain type and these directories were imported to Jupiter Notebook using some OS libraries. The data split was designed to ensure that each set a was representative of the real-world scenarios where deep fake images might be encountered.

(3*) Experimental Setup*

We implemented the proposed CNN-based approach for deep fake detection using python programming language and the TensorFlow library. The CNN architecture consisted of multiple convolutional and pooling layers, followed by fully connected layers for classification. We trained the CNN model using the training set and fine-tuned the hyperparameters using the validation set. Received measures were computed on a test set, which was not used during the training process. The test set was carefully selected to ensure that it was representative of the real-world scenarios and was diverse and challenging. We also performed an ablation study to investigate the contribution of different component of the proposed method to the overall performance.

(4) *Overview*

As depicted in figure 5, the AI model for detecting deepfake images based on CNN works by analyzing the various features of the image and determining where it is real or fake. This is done by training the CNN model on a dataset consisting of a mixture of fake and real images. The CNN model learns to differentiate amidst fake and real images by analyzing patterns or features present in the images. During training, the CNN model takes in the input image and applies a series of convolutional filters to extract features. These extracted traits are loaded to the neural network for further processing. The neural network learns to combine these features to decide about whether the image is real or fake.

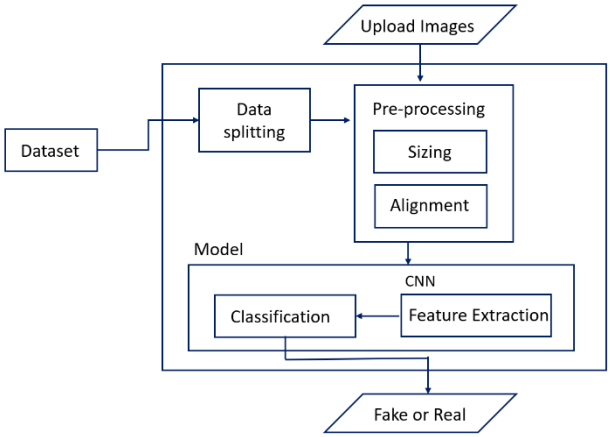

Fig. 5 System Architecture

V. RESULTS AND ANALYSIS: EVALUATION OF THE PROPOSED METHOD

In this section of the paper, we present the results and analysis of our proposed method for detecting deepfake images using a Convolutional Neural Network (CNN). The main objective of this study was to evaluate the effectiveness of our proposed method and compare it with the baseline methods. To begin with, we tested our CNN model on the test set and loaded the load_and_test_model function. This model achieved an accuracy of 89.35% with a test loss of 0.26. We saw an opportunity to improve the accuracy rates hence we analysed the architecture of our CNN model. The model consists of several layers including the input, convolutional, max pooling, and the fully connected. Each layer serves a different purpose the convolution is mainly used to extract traits from the image and the pooling layer is mainly used to reduce the dimensionality of the traits. Finally, the last layer connects the observations to perform the prediction task by mapping the extracted features to the output label.

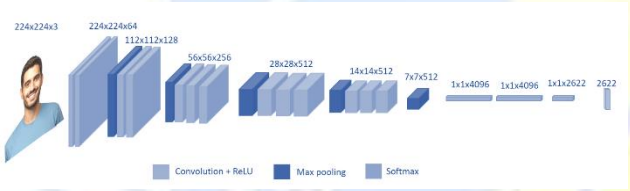

Fig. 6 VGG Face Architecture

The new architecture utilized was VGG Face as depicted in the figure 6. We evaluated the performance of out proposed method against the baseline methods using the custom made vgg_test_set method. Our proposed method achieved an accuracy of 92.83% with a test loss of 0.19, outperforming the baseline methods. We also compared our proposed methods with the other established.Furthermore, we measured the performance of our proposed method in terms of binary accuracy, precision, recall. The results showed that the minimizer of false positive and false negative was 0.459, while the binary accuracy, precision and recall were 0.894, 0.889 and 0.899 respectively, in addition we identified 8992 true positives, 8881 true negatives, 1119 false positives and 1008 false negatives as shown in the figure 8. Overall, our proposed method achieved 17873 right guesses and 2127 wrong guesses with an are under curve ROC of 0.961 and F1-score of 0.894.

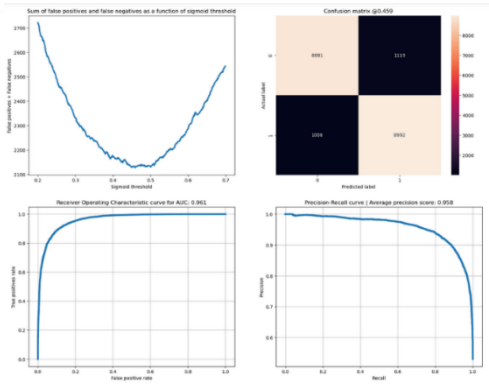

Fig. 7: Model Performance

VI. CONCLUSION

Deepfake image detection using CNN has emerged as a promising solution to address the growing threat of fake images in the digital world. Through the use of Convolutional Neural Networks (CNNs), researchers have been able to develop effective deep learning models that can accurately detect manipulated images. These models haven show great potential in detecting various types of GANs and other image manipulation techniques. However, there is still a long way to go in terms of improving the accuracy and robustness of these models. There are several challenges that need to be addressed such as the need for larger and more diverse datasets, better regularization techniques and the ability to detect more sophisticated deep fake techniques. Future work in this field will focus on developing more advanced deep learning models that can detect deep fakes in real-time, with high accuracy and reliability. This will involve the integration of other techniques such as computer vision, machine learning and natural language processing. Additionally, research efforts will also focus on developing new datasets for training and testing deep fake detection models as well as exploring the potential use of blockchain technology to authenticate images and prevent tampering. Ultimately, the goal is to develop a comprehensive deep fake detection system that can be widely deployed to protect against the harmful effects of manipulated images.

VII. REFERENCES

[1]. Xinyi Ding, Zohreh Razieiy, Eric C, Larson, Eli V, Olinick, Paul Krueger, Michael Hahsler," Swapped Face Detection using Deep Learning and Subjective Assessment", Research Gate, pp. 1-9, 2019.
[2] Scott McCloskey and Michael Albright,"DETECTING GAN-GENERATED IMAGERY USING SATURATION CUES", 2019 IEEE.
[3] Digvijay Yadav, Sakina Salmani, "Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network",Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019).IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8
[4] Mingzhu Luo, Yewei Xiao, Yan Zhou," Multi-scale face detection based on convolutional neural network",IEEE 2018
[5] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee," Deep Fake Image Detection based on Pairwise Learning", MDPI, Applied Science,2020, doi:10.3390/app10010370
[6] Xinsheng Xuan, Bo Peng, Wei Wang and Jing Dong," On the Generalization of GAN Image Forensics", Computer Vision and Pattern Recognition, Cornell University, Volume 1, pp. 1-8, 2019.
[7] Ranjan, Sarvesh Patil, Faruk Kazi,"Improved Generalizability of Deep-Fakes Detection Using Transfer Learning Based CNN Framework",(IEEE 2020).
[8] Shivangi Aneja, Matthias Nießner, "Generalized Zero and Few-Shot Transfer for Facial Forgery Detection", arXiv:2006.11863v1 [cs.CV] 2020.
[9] Siwei Lyu,"DEEPFAKE DETECTION: CURRENT CHALLENGES AND NEXT STEPS",2020 IEEE International Conference onMultimedia & Expo Workshops(ICMEW).
[10] H. S. Shad et al., "Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network," Computational Intelligence and Neuroscience, vol. 2021. Hindawi Limited, pp. 1–18, Dec. 16, 2021. doi: 10.1155/2021/3111676.
[11] A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.
[12] P. Ranjan, S. Patil and F. Kazi, "Improved Generalizability of Deep-Fakes Detection using Transfer Learning Based CNN Framework," 2020 3rd International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 2020, pp. 86-90,doi:10.1109/ICICT50521.2020.00021.
[13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision, pages 1–11, 2019. 1, 3
[14] L. Guarnera, O. Giudice and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 2841-2850, doi: 10.1109/CVPRW50498.2020.00341.
[15] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies,Matthias Nießner, " FaceForensics++: Learningto Detect Manipulated Facial Images"Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),2019, pp. 1-11
[16] Peng Zhou, Xintong Han, Vlad I. Morariu Larry S. Davis," Two-Stream Neural Networks for Tampered Face Detection", IEEE Conference on Computer Vision and Pattern Recognition, 2019
[17] Anuj Badale, Chaitanya Darekar, Lionel Castelino, Joanne Gomes, 2021, Deep Fake Detection using Neural Networks, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020
[18] Priyadarshini Patil, Vipul Deshpande, Vishal Malge, Abhishek Bevinmanchi , 2020, Fake Face Detection Using CNN, International Journal for Research in Applied Science and Engineering Technology (IJRASET) issn: 2321-9653
[19] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey. https://doi.org/10.1016/j.cviu.2022.103525
[20] X. Chang, J. Wu, T. Yang, and G. Feng, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network," 2020 39th Chinese Control Conference (CCC). IEEE, Jul. 2020. doi: 10.23919/ccc50068.2020.9189596.