**AI Against AI: A Deep Learning Approach to Detecting Deepfakes**

**Mayank Narain**

**Department of Computer Science and Application, Sharda University, Greater Noida, Uttar Pradesh, India**

**2022471176.mayank@ug.sharda.ac.in**

**ABSTRACT**

The term "deepfake technology" describes the process of manipulating or editing multimedia information, such as photos or videos, using machine learning and artificial intelligence to produce fake however incredibly lifelike portraits of actual people. Deepfake material is becoming more and more common, which has sparked worries about its possible misuse for identity theft, extortion, and the dissemination of false information. In reaction to this new danger, scientists have created several tools and strategies for identifying deepfake. Convolutional Neural Networks are among the most exciting method for deepfake detection. Convolutional Neural Networks are a strong option for deepfake detection because of their shown ability to analyze and recognize patterns in photos. Deep learning methods have been widely applied in computer vision recently, particularly in the areas of picture editing and detection. The development of deepfakes, which are produced by deep learning algorithms and may be very challenging to distinguish from authentic photos, has also been made easier by these developments. Researchers have suggested using deep learning algorithms—more especially, Convolutional Neural Networks—for deepfake detection in order to overcome this difficulty. These networks are able to understand the distinctive traits and patterns that separate legitimate material from deepfake because they are trained on vast datasets of real and fake pictures.

Keywords : - Deep learning, Convolutional neural network, Artificial Intellige

**INTRODUCTION**

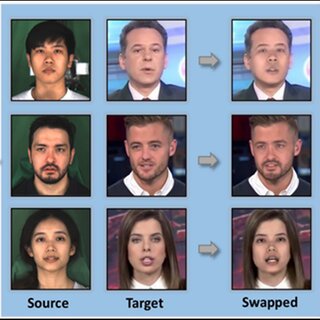
As technologies growing more readily available to all users, several deepfake films have proliferated over the internet. Deepfake relates to modified online content, such as photos or videos, in which the visual or audio representation of one person is replaced with another person's resemblance. As a matter of fact, deepfake is becoming one of the most severe challenges in modern civilization. Deepfake is often used to fraudulently obtain the faces of noticeable Hollywood celebrities over pornographic picture films. Deepfake was also used to spread false information and rumors among politicians [1]. These dangerous applications of deepfakes can have major consequences for our society, including the propagation of false information, particularly on the internet.

According to recent studies, deepfake pictures and videos are widely distributed over digital platforms. Detecting deepfake videos and pictures therefore has grown increasingly essential and essential. To motivate researchers, numerous companies, such as the United States Defense Advanced Research Projects Agency (DARPA), Facebook Inc, and Google started a research program in seeking to identify and prevent deepfakes. [2]. As a result, numerous deep learning algorithms, including long short-term memory (LSTM), recurrent neural network (RNN), and hybrid approaches, have been suggested to detect deepfakes in photos and videos, as well as to stimulate further study in this field [3]. According to recent research, deep neural networks do exceptionally well in detecting false reports and gossip in internet.

This work primarily focuses on providing a comprehensive study for deepfake detection using deep-learning methods such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long short-term memory (LSTM). This survey will be useful and beneficial for researchers in this field as it will give: 1) details summary of the current research studies; 2) datasets used in this field; 3) the limitations of the current approaches and insights of future work.

As therefore, detection technologies have to modify in order to maintain track with such technologies. Among the most challenging aspects of recognizing deepfakes knows those that are highly believable. These deepfakes may be constructed using powerful AI algorithms, rendering them practically hard to identify from actual material. This implies that even the most sophisticated detection methods may be unable to recognize them. To summarize, deepfakes technology is capable to be exploited maliciously, hence efficient identification algorithms must be implemented. While there are numerous various strategies for detecting deepfakes, each has drawbacks [4]. As therefore, detection technologies have to modify in order to maintain track with such technologies. Among the most challenging aspects of recognizing deepfakes knows those that are highly believable. These deepfakes may be constructed using powerful AI algorithms, rendering them practically hard to identify from actual material. This implies that even the most sophisticated detection methods may be unable to recognize them. To summarize, deepfake technology is capable to be exploited maliciously, hence efficient identification algorithms must be implemented. While there are numerous various strategies for detecting deepfakes, each has drawbacks. Continuous investigation and improvement in algorithms for identification will be critical for keeping ahead of the latest developments in deepfake technology.

* 1. **DEEPFAKE**

The term "deepfakes," which combines the terms "deep learning" and "fake," refers to a specific type of video that is produced by using facial image superimposition methods to a source person's video. The result is a video that shows the target person acting or speaking in the same way as the source person [5]. This belongs to the face swap genre of deepfakes. Deepfakes are, essentially, artificial intelligence-generated material that also fits into lip-sync and puppet-master categories. Videos that have been altered to match the mouth movements with an audio recording are called lip-sync deepfakes as shown in Fig.1.1.

Videos of a target individual (puppet) that is animated to mimic the facial expressions, eye movements, and head movements of another person (master) seated in front of a camera are known as puppet-master deepfakes.

Fig.1 1: deepfake generated images

* 1. **HISTORY OF DEEPFAKE**

The idea of deepfakes originated in the 1990s when scientists started utilizing computer-generated imagery (CGI) to produce lifelike human pictures. A Reddit user who formed a forum for the exchange of deepfake pornography made with open-source face-swapping technology is credited with coining the term "deepfake" in 2017.

The next generation of sophisticated picture, video, and audio deepfakes was made possible in 2014 by Ian Goodfellow and his team's breakthrough in deep learning, which led to the introduction of Generative Adversarial Networks (GANs) [6]. The decade of the 2010s saw significant developments in deepfake technology because to the availability of massive datasets, advances in machine learning, and increased processing power.

With the increased accessibility of deepfake generating tools, amateurs and hobbyists started honing them for both malevolent and entertaining reasons, such as producing deepfake pornography. The democratization of strong tools brought attention to the necessity of having sufficient deepfake detection techniques [6].

Experts voiced concerns in 2018 over the implications of deepfake technology's rapid advancement. Large internet companies started enforcing rules to restrict the usage of deepfakes on their networks. In 2019, a number of nations, the US included, investigated passing laws to control the production and dissemination of deepfakes [7].

* 1. **DEEPFAKE CREATION**

Deepfakes are produced by utilizing deep learning algorithms to replace a target's face in a picture or video with another person's. Designers and online communities refined this method to produce remarkably simple-to-operate programs that are easily accessible online, such as Face Swap and FakeApp [8]. An auto encoder-decoder network is the foundation of deepfake. Encoders, which are commonly used in image compression, depend on deep neural networks (DNNs), and when an error in the network is introduced, this determines the network to output the initial input in a compressed format [9]. High-quality picture compression is made practical by progressively advanced encoders, which can make deep fake tasks easier to complete because less processing power is needed. Two auto encoders are trained to create deepfakes.

The target image is generated using the source image's decoder, which will produce an image of the target with features of the source after one auto encoder learns the features of the source image and the second encoder learns the features of the target image [10]. The two encoders then share their parameters to generate the deep fake image as shown in Fig. 1.2

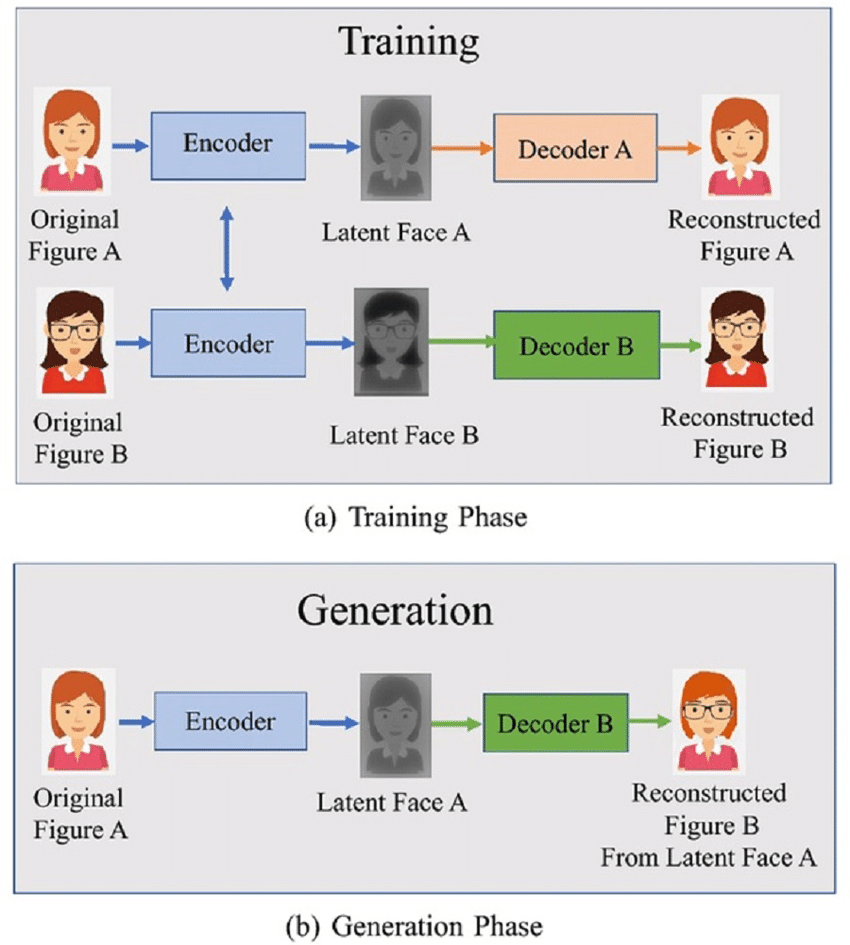


Fig.1.2: Deepfake creation layout

The range of data that can be used to train a deep neural network determines how well this approach performs. Fortunately, social media platforms like Instagram and YouTube have an extensive collection of publicly accessible images and videos, particularly those featuring public figures and celebrities. As a result, these individuals are frequently the targets of deepfakes and suffer the most from them [11]. There are several datasets available from scientific organizations like Celeb- DF for usage in deepfake.

These ideas may be utilized to create deepfake photos and/or movies. Obviously, pictures can be created more quickly than films because they require less processing and are smaller in size. The threat of deepfake is evident given the development of deepfake tools and the growth in fake news. Hard-to-detect fake images are getting simpler to create using user-friendly deepfake tools, and fake news that may employ these images is widely disseminated online.

* 1. **CNNs (Convolutional neural network)**

A category of Deep Learning neural network design that is frequently utilized in computer vision is the Convolutional neural network (CNN). The branch of artificial intelligence known as "computer vision" gives computers the ability to comprehend and analyze images and other visual input [12].

Three different sorts of layers are present in a typical neural network.

A) Input layer: This layer is where we provide data into our model. The entire number of features in our data (or, in the case of a picture, the number of pixels) is equivalent to the number of neurons in this layer.

B) Hidden Layer: The hidden layer receives the input that was previously given into the input layer. Several hidden layers may exist, based on the amount of the data and our model. The number of neurons in each hidden layer varies, although they are usually more than the number of characteristics [13]. The network is non-linear because each layer's output is calculated by multiplying the output of the layer before it by the adjustable weights of that layer, adding learnable biases, and then applying the activation function.

C) Output Layer: The probability score for each class is obtained by feeding the output of the hidden layer into a logistic function such as the sigmoid or softmax, as shown in Fig.1.3.

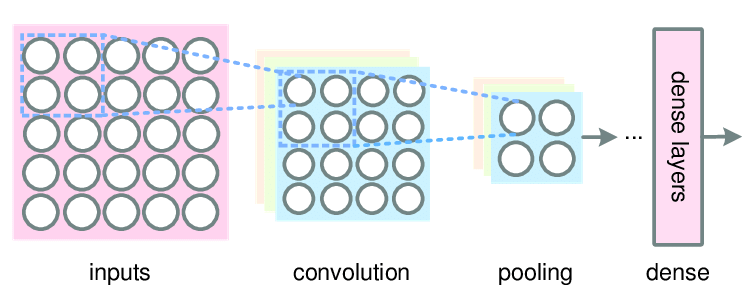


Fig.1. 3: Convolutional Neural Network

* 1. **MesoNet**

A Convolutional neural network (CNNs) called MesoNet was created particularly to identify deepfakes. Deepfake videos appear frequently on social media because of the low-quality, compressed nature of these videos, which makes microscopic analysis based on image noise impractical. MesoNet also believes that detecting deep fakes at a higher semantic level is challenging because even humans have difficulty spotting deep fakes occasionally. As a result, MesoNet uses an intermediary method that involves a deep neural network with few layers. This network has four layers that are composed of pooling and sequential convolutions in the initial stages, and then a dense network with a single hidden layer. When extracting features from an image, convolution, and pooling are often employed in that order: the convolution layer separates the features, and the pooling layer produces a down-sampled form of the feature mapping Fig.1.4.

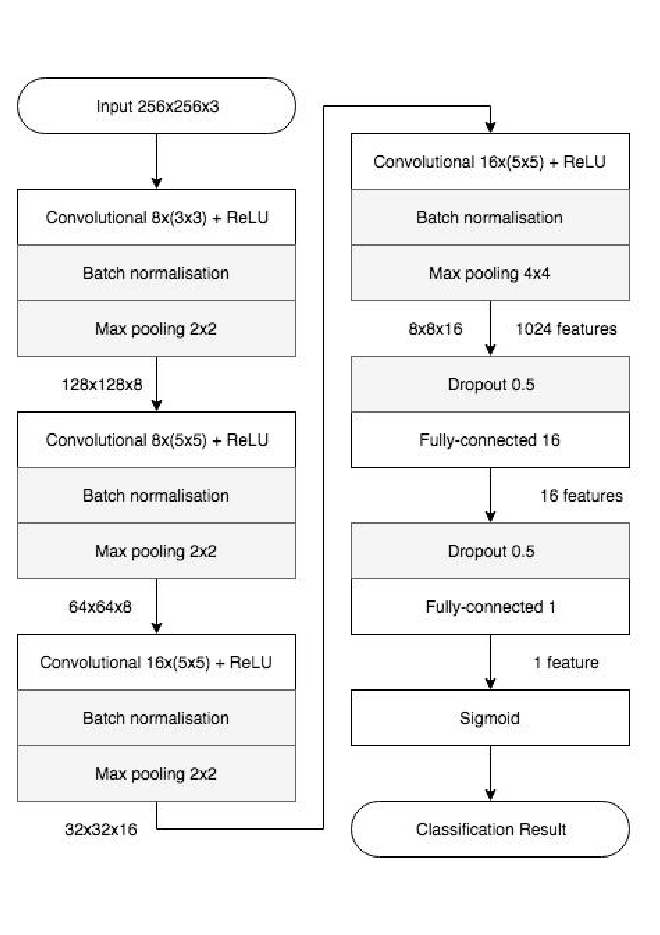


Fig.1.4: MesoNet’s architecture includes four layer of CNNs layer with pooling layer

The fully connected layers apply Dropout to regulate and strengthen their resilience, while the Convolutional layers use ReLU activation functions to add non-linearities and Batch Normalization [14] to regularize the output to increase generalization.

**REVIEW OF LITERATURE**

**2.1. MesoNet**

*Rebello, L .et al., 2023* in a significant literature have quoted that MesoNet is a neural network built primarily to identify deep fakes, but it would also be used for other purposes. MesoNet manages the noise produced by low-quality video processing, which impedes analysis. Deepfakes jeopardizes facial recognition and internet content. This deception is risky and can be exploited to impersonate a legitimate user. Our approach will propose a temporal-aware method for automatically detecting deepfake video [15]

*Aduwala, S. A,* *et al, 2021* Researcher has reported that using MesoNet as a baseline, GAN was trained and extracted discriminator as a dedicated module to detect Deepfakes. Researcher has also tested several discriminator architectures using multiple datasets to explore how the efficacy of the discriminator varies with different setups and training methods. Finally, proposed a model to boost the efficacy of a group of GAN discriminators using ensemble methods. Results show that GAN discriminators, even augmented by ensemble methods, do not perform well on videos from unknown sources [16].

*Afchar.D et al., 2018* has explained this paper which has reported that the system detects face tampering in films automatically and effectively, with an emphasis on two contemporary methods—Deepfake and Face2Face—that are used to create forged videos that are incredibly convincing. Videos often don't lend themselves well to standard picture forensics techniques because of the high data degradation caused by compression. To focus on the mesoscopic qualities of pictures, this study offers two networks that both have a limited number of layers and use a deep learning methodology. We test such rapid networks on an already-existing dataset as well as one we created from internet videos. With over 98% and 95% detection rates, respectively, for Deepfake and Face2Face, the tests show an extremely successful detection rate [17].

*Xia. Z et al., 2022* have published this paper which states that MesoNet-based Deepfake video detection technique with a preprocessing module. Initially, the preprocessing module is set up to enhance the identification between multi-color channels by preparing the cropped face photos. The traditional MesoNet is then given the preprocessed photos. The suggested method's detection performance is confirmed on two datasets; it outperforms the existing approaches with an AUC of 0.974 on FaceForensics++ and 0.943 on Celeb-DF. More significantly, the detection rate may still exceed 88% even under conditions of severe compression [18].

*Gupta.A et al*., 2022 in the significant literature have quotes that the study adds to the current discourse on deepfake detection by evaluating the MesoNet model's efficacy in particular. It employs the Face2Face deepfake dataset, which is renowned for its skill at facial recreation, and concentrates on the study of facial micro-expressions. Researcher also explained that the study aims to assess MesoNet effectiveness by analyzing its performance in several dimensions, optimizing the model to get better outcomes, and acquiring a sophisticated understanding of its capabilities. The results show a significant improvement, with MesoNet exceeding the prior accuracy of 89.1% with an accuracy of 90.4%. Extensive outcomes after meticulous adjustments to regularization parameters and activation functions highlight the importance of hyper parameter optimization in deep learning models [19].

**2.2. CNNs based Algorithm**

*Li, Z. et al 2020, Yang, W. et al 2020,* in this significant literature have quoted that the Convolutional neural networks are capable of extracting features from data that has convolution patterns. CNN extracts features automatically, in contrast to conventional feature extraction techniques. Visual perception serves as an inspiration for CNN's architecture. Artificial neurons are analogous to biological neurons; CNN kernels mirror distinct receptors capable of reacting to diverse characteristics; activation functions mimic the process by which neural electric signals over a specific threshold are forwarded to the subsequent neuron. To train the CNN system as a whole to learn what researcher has intended, people devised loss functions and optimizers.  
When compared to other artificial neural networks, CNN has several benefits. 1) Local interactions. 2) Weight sharing: This reduces parameters and speeds up convergence by connecting each neuron to a smaller set of neurons instead of all the neurons in the preceding layer. Further reduction of parameters might occur when a set of links have the same weights. 3) Dimensionality decreases by down sampling using the idea of image local connection, a pooling layer down samples a picture. This can cut down on data while keeping important information. It can also cut down on parameters by eliminating elements that are not necessary. CNN is one of the finest representative algorithms in the domain of deep learning because of these three attractive features. Specifically, four parts are usually required to construct a CNN model. A crucial stage in the feature extraction process is convolution. Convolutional outputs are referred to as feature maps. A convolution kernel set to a specific size will result in information loss at the boundary. Therefore, padding is added to the input to make it larger using a zero value, which might indirectly change the size. Additionally, stride is utilized to regulate the convolving density. The density decreases with increasing stride length. Feature maps, which are prone to over fitting issues, are composed of numerous features following convolution Because of this, pooling—also known as down-sampling—including average and max pooling—is suggested as a way to eliminate repetition[20].

*Jogin, M. et al., 2018 & Mohana, M. et al.,2018* has explained this paper which discussed the architectures of some classic networks, but there are many new-generation networks, after their work, have been proposed, such as Mobile Net v3, Inception v4, and ShuffleNet series, which deserve researchers’ attention. Besides, the work reviewed applications of CNN for object detection [21].

*Huang et al.,* 2018has explained this paper which has presented a new scaling technique that uses a straightforward yet extremely potent compound coefficient to scale all three dimensions of depth, breadth, and resolution equally. Researcher has used MobileNets and ResNet to show how effective this approach is. We create a new baseline network via neural architecture search, then scale it up to produce a family of models known as EfficientNets that outperform earlier ConvNets in terms of accuracy and efficiency. Specifically, EfficientNet-B7 outperforms the best-performing ConvNet by 8.4 x and 6.1 x on inference, and achieves state-of-the-art 84.4% top-1/97.1% top-5 accuracy on ImageNet [22].

Shi, W et al., 2019 has explained this paper which suggested a Convolutional neural network (CSNet) for an image CS framework. CSNet consists of a sampling network and a reconstruction network that are tuned in tandem. In order to improve reconstruction, the CS measures retain more structural information from the training pictures thanks to the sampling network's adaptive learning of the sample matrix.The {0,1} binary matrix, the {-1,+1} bipolar matrix, and the floating-point matrix are the three types of sample matrices that are learnt. The final two matrices are specifically made to be easily implemented on hardware and stored. The reconstructed network learns an end-to-end mapping between the CS measurements and the rebuilt pictures. It consists of two parts: a linear initial reconstruction network and a non-linear deep reconstruction network [23].

**2.3. GNNs base algorithm**

*Huang, Y. et al., 2022* has published this paper which provided a fine-grained atomic operations-based inheritance, crossover, and mutation operator search technique for evolutionary graph neural network architectures. In particular, researcher also created two new overlap operators, LayerCross and GNNCross, at various granularity levels. Also, conducted Tests on three distinct graph learning tasks reveal that the neural networks produced by our approach function on track with the manually designed and automatically generated baseline GNN models [24].

*Pang L. et al. 2022* has explained this paper which developed Smart Library's clever data extraction system Compared to earlier broad approaches, the graph neural network (GNN) algorithm yields 80% more results. When it comes to categorized nodes, link estimation, node grouping, or network graphical representation, graph neural networks are a more beneficial approach that may significantly increase the effectiveness of information extraction [25].

Ji, Y. et al., 2021 in a this significant literature have created a brand-new detection of anomalies technique using data from many sources. Initially, researcher utilizes the method of spectral grouping approach to extract features and combine data from several sources and Second, they executed a fine-gained anomalous social event detection, disclosing the dangerous occurrences and ensuring the security of key systems by utilizing the power of deep graph neural networks (Deep-GNN). The outcomes of our experiments suggest that our framework works better than baseline anomalous event detection techniques and has excellent resilience, stability, and monitoring precision [26].

Hu, Z., et al 2022 has explained this paper which suggested a smart monitoring technique based on the graph neural network (GNN) algorithm for the evolution of rural settlement morphology. First, a detailed explanation of the graph neural network (GNN) algorithm-based picture feature extraction, analysis, and processing procedure was given. Second, the graphical neural network algorithm was utilized to successfully extract the morphological features of rural neighborhoods in conjunction with the changing traits of rural community morphology evolution and scale development. The feature was used for collecting data for tracking and characterizing the dynamic changes in rural settlement morphology and scale. Finally, immediate time monitoring of the development of rural settlement morphology was achieved through tests and the use of the graph neural network method [27].

**2.4. Other Techniques**

*Aghasanli, A et al.,2023* has established this paper to quote that deep learning algorithms that produce less fuzzy and more detailed photos include Stable Diffusion. Researcher created a deepfake detection method to differentiate between authentic and synthetic pictures produced by different Diffusion Models. The deepfake detection approach leverages characteristics from optimized Vision Transformers (ViTs) in conjunction with established classifiers like Support Vector Machines (SVM). Researcher also analyzed the support vectors of the SVMs to show the interpretability-through-prototypes capability of the suggested technique [28]

.

*Rana,M, S. et al.,2020* have published this paper which provided the DeepfakeStack deep ensemble learning method for identifying these kinds of manipulated vedios. By merging several state-of-the-art DL-based classification models which suggested method produced for better composite classifier. Researcher also demonstrate that DeepfakeStack works better than other classifiers and detecting Deepfake with an accuracy of 99.65% and an AUROC of 1.0 score. Thus, this approach offers a strong foundation for developing a Realtime Deepfake detector [29]

.

*Vashishtha, S. et al., 2022* have published this paper which created a deep learning ensemble system capable of distinguishing between authentic and counterfeit photos. A unique strategy is given to extract the apparent motion of picture pixels using the suggested optical flow technique model- OptiFake, yielding more accurate findings than current state-of-the-art methods. The extraction techniques and ensemble model are tested on the FaceForensics + + dataset, which yielded accuracy of 86.02% for the DeepFake subset and 85.7% for the FaceSwap subset [30].

*Heidari, A. et al 2024* in this significant literature have offered a novel approach that preserves data source privacy by utilizing blockchain-based federated learning (FL). SegCaps and Convolutional neural network (CNN) techniques are used to increase visual feature extraction, and then capsule network (CN) training is applied to improve generalization. A new method of data normalization is presented to address variability in data originating from several worldwide data sources. To improve DL performance, preprocessing techniques and transfer learning (TL) are applied. Researcher also conducted tests to carefully verify and confirm the efficiency of a method [31].

*Hubálovský, Š. Et al 2022* has explained this paper which purposed a method for identifying manipulated videos called the Only Look Once-Local Binary Pattern Histogram (YOLO-LBPH). To identify a face in an image or a frame of a video, YOLO is utilized. Using an EfficientNet-B5 technique, the spatial features are retrieved from the facial picture. The Local Binary Pattern Histogram uses spatial feature extractions as input to obtain temporal data [32].

*Agarwal, A. et al 2021* have published this paper which introduced a unique method for merging features in the frequency and spatial domains to derive a common selective model for deepfake classification, called MD-CSDNetwork. The MD-CSDNetwork is a new type of cross-stitched network that carries frequency and spatial data in two parallel arms. Researcher also reported that these streams of multi-domain input data may be seen as connected supervision signals that guarantee improved generalization and performance [33].

*K, V. et al 2023,& Trojovský, P. et al 2023*, have publied this paper which purposed a VIOLA Jones (VJ) method which suggested as a means of choosing the optimal features for the Capsule Graph Neural Network (CN). Capsule-based node feature extraction is a technique used to enhance the performance of graph neural networks. CelebDF-FaceForencics++ (c23) datasets, which integrate FaceForencies++ (c23) and Celeb-DF, are used to assess the experiment [34].

**2.5. Future Aspects**

*Farid. H et al.,2010* has explained this paper which investigated forensic methods for identifying digital media manipulation, setting the stage for further research. These early techniques mostly depended on statistical modeling and pixel-level analysis, which, although useful at the time, proved difficult to use with subsequent, more complex modifications [35].

*Xu. et al.,2016* in a significant literature have quoted that presented convolutional neural networks (CNNs), which showed considerable accuracy gains over previous techniques, for the identification of digital forgeries [36].

*Rössler.A et al.,2017* have published this paper which created FaceForensics, an extensive dataset that made it possible to train deeper learning models for deepfake detection that were more reliable[37].

*Li.et al., 2019*has explained this paper which introducedGAN-based deepfake detection model because deepfakes could create extremely realistic-looking fake material, they presented new issues. Created innovative deep learning architectures that enhanced detection accuracy by utilizing both temporal and spatial aspects of films. Suggested using recurrent neural networks (RNNs) to identify small manipulations by capturing variations in face emotions and movements over time [38].

*Qi et al., 2020* in this significant literature have highlighted the value of using a variety of data sources to identify deepfakes, while using only visual data to identify sophisticated fakes was shown to be insufficient. Real-time detection methods (Nguyen et al.,2020), were also introduced during this period. They used lightweight DL models to produce quick and accurate findings, making them appropriate for use in social media platforms and streaming services [39]

**MATERIALS AND METHODOLOGY**

To develop a neural network model, MesoNet/Meso-4, for deepfake detection, requires the following material and tool:

**HARDWARE REQUIREMENTs**

1. Computer System with high performance CPU with sufficient storage space (at least 16 GB) for storing dataset and model checkpoints.
2. High-performance GPU (preferred NVIDIA RTX 3080) with at least 8 GB VRAM

**SOFTWARE REQUIREMENTS**

1. Operating system : window 10
2. Python environment: Python 6.0
3. Python Libraries:
4. PyTorch for deep learning framework
5. NumPy for numerical operations
6. OpenCV for image processing
7. Scikit-learn for additional machine learning tools
8. Matplotlib or Seaborn for data visualization
9. Pandas for data manipulation
10. Development Tools: Integrated Development Environment (IDE) VS Code, or Jupyter Notebook.
11. Git for version control.

**DATASET**

1. Deepfake dataset : dataset of both real and fake images/vedios
   1. FaceForensics++: dataset containing real and manipulated videos.
   2. DFDC (Deepfake Detection Challenge): A large dataset provided by Facebook.
   3. Celeb-DF: A dataset with high-quality deepfake videos.
2. Processing tool: scripts to process dataset

**MESONET MODEL COMPONENTS**

1. Network architecture

This network initiates with a series of four layers composed of convolutions and pooling layers then these layers are followed by a dense network containing one hidden layer. To enhance the network's ability to generalize, ReLU activation functions are employed in the convolutional layers, as shown in fig. 3.1. These functions introduce non-linearities and batch normalization techniques, which serve to regularize the output and prevent the occurrence of vanishing gradient effect. Additionally, Dropout is utilized in the fully connected layers to further improve their robustness and promote regularization

|  |
| --- |
| Convolutional +ReLU |
| Batch normalisation |
| Max pooling |

|  |
| --- |
| Convolutional +ReLU |
| Batch normalization |
| Max pooling |

|  |  |
| --- | --- |
| Convolutional+ReLU |  |
| Batch normalization |  |
| Max pooling |  |

|  |
| --- |
| Convolutional +ReLU |
| Batch normalisation |
| Max pooling |

|  |
| --- |
| Dropout |
| Fully connected |

|  |
| --- |
| Dropout |
| Fully connected |

|  |
| --- |
| Sigmoid |

Fig.3.1: MesoNet model architecture

1. Training and Validation Code: To train the model on the dataset used script and to evaluate the model process used Validation.
2. Optimizer : Adam

**4.2 Results**

MesoNet performed on particular images predicted output is between 0- 1.

According to their corresponding degree of confidence in the prediction.

* 1. Observation 1 prediction likelihood is 0.1297 which corresponding to the high degree of confidence in the prediction which means image is deepfake
  2. Observation 2 images data sets prediction likelihood are above 0.5 mainly 0.9 which represent that those images data set are not manipulated
  3. Observation 3 images data set prediction likelihood are between 0.2 – 0.4 which means those images are slightly manipulated
  4. Observation 4 images data set prediction likelihood are between 0.01-0.1 which those images are deepfake
  5. Observation 5 images data set out prediction are between 0.01- 0.3 which means those images are misclassified deepfake

**CONCLUSION**

The Convolutional neural network (CNNs), MesoNet model represents a significant advancement in the field of facial video forgery detection, offering a compact and efficient solution tailored for deepfake detection. Through its lightweight architecture, MesoNet or Meso-4 model effectively balances performance and computational efficiency, making it suitable for real-time applications. For increased efficacy, the MesoNet model leverages deep learning to detect subtle manipulations within facial videos, enabling accurate deepfake identification. Its innovative design optimizes resource utilization, enhancing its ability to process large volumes of data efficiently. This pioneering approach in deepfake detection enhances the security and integrity of multimedia content in various applications, catering to the evolving demands of modern digital ecosystems. The model's design, incorporating four Convolutional layers with ReLU and pooling operations followed by dense layer and one hidden layer, enables it to capture intricate facial features and predict subtle manipulations characteristic of deepfakes up to 92% into four category which are correct real, correct deepfake, misclassified real and misclassified deepfake. This innovative approach not only enhances the accuracy of forgery detection but also reduces the resource overhead typically associated with deep learning models. Consequently, MesoNet stands out as a practical and powerful tool for combating the growing threat of deepfake technology, providing a robust framework for safeguarding digital media integrity.

**5.1 Validation**

In our work, we validated the efficacy of the MesoNet architecture, a specialized neural network designed to detect facial video forgeries. The MesoNet model is notable for its compact structure, making it efficient for deployment in environments with limited computational resources. Our validation process involved rigorous testing on benchmark datasets, including celeb- DF, YouTube testing videos and the Deepfake Detection Challenge (DFDC) dataset. These datasets contain a diverse array of real and manipulated videos, providing a comprehensive assessment of the model's performance.

We implemented the MesoNet following the architecture outlined by Afchar et al., ensuring faithful replication of the original design. The model comprises multiple Convolutional layers that extract critical features from facial images, followed by dense layers that facilitate classification. Our training regimen included extensive data augmentation techniques to enhance the model's robustness against various forms of video manipulations.

During validation, we evaluated the model using standard metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). The results demonstrated that MesoNet effectively distinguishes between authentic and forged videos with high accuracy up to 92%, maintaining competitive performance compared to more complex models. Additionally, its relatively low computational requirements underscore its suitability for real-time applications, making it a valuable tool in the fight against deepfake dissemination.

Overall, our findings confirm that MesoNet is a powerful and efficient model for deepfake detection, capable of delivering reliable results without the need for extensive computational resources.

**5.2 limitations**

One limitation of the MesoNet model, as applied to facial video forgery detection, lies in its susceptibility to over fitting when trained on limited or non-diverse datasets. Despite its compact architecture being beneficial for efficiency and quick deployment, this simplicity can also hinder its ability to generalize across a wide variety of deepfake techniques and video qualities. Consequently, MesoNet may struggle to accurately detect forgeries in videos that exhibit variations significantly different from those in the training data, reducing its robustness and reliability in real-world applications where deepfakes continue to evolve rapidly.

**5.3 Future aspect**

The future development of MesoNet holds promising potential for enhancing the accuracy and efficiency of deepfake detection . One of the primary avenues for advancement lies in the integration of more sophisticated and diverse datasets, which can help the model generalize better to various types of forgeries. Additionally, incorporating techniques such as transfer learning and domain adaptation could enable MesoNet to leverage pre-existing knowledge and adapt to new, unseen types of deepfake manipulations more effectively.

Moreover, future iterations of MesoNet could benefit from advancements in neural network architectures, such as the integration of attention mechanisms or transformer models, which have shown considerable success in various computer vision tasks. These enhancements could provide the model with a better understanding of contextual information, leading to improved detection performance [40].

Another exciting prospect is the deployment of MesoNet in real-time applications. Optimizing the model for faster inference on edge devices or incorporating it into real-time video processing pipelines could make it a valuable tool for content moderation on social media platforms and in forensic investigations. Additionally, ongoing research into explainable AI could be integrated into MesoNet to provide transparency in the decision-making process, thereby increasing trust and adoption in critical applications [41].

Overall, the continuous evolution of computational power, combined with innovative research in machine learning, presents a robust pathway for MesoNet to become an even more effective and widely adopted solution for combating the growing challenge of deepfake videos.

REFERENCES

1. Chien, J.-T. (2020). Deep Bayesian multimedia learning. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)* (pp. 3).
2. Khalil, H. A., & Maged, S. A. (2021). Deepfakes Creation and Detection Using Deep Learning. 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 4).
3. Nataraj, L., Mohammed, T., Manjunath, B. S., Chandrasekaran, S., & Flenner, A. (2019). Detecting GAN generated fake images using co-occurrence matrices. *Electronic Imaging, 2019*, 532-1-532-7.

1. Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8695-8704).
2. Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. *Journal of Cybersecurity and Privacy*, *2*(1), 89-106.
3. Ahmed, S. R., Sonuç, E., Ahmed, M. R., & Duru, A. D. (2022, June). Analysis survey on deepfake detection and recognition with convolutional neural networks. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-7). IEEE.
4. Deshmukh, A., & Wankhade, S. B. (2020). Deepfake detection approaches using deep learning: a systematic review. *Intelligent computing and networking: proceedings of IC-ICN 2020*, 293-302.
5. Diakopoulos, N. and Johnson, D. (2020). Anticipating and addressing the ethical implications of deepfakes in the context of elections. New Media & Society, 23(7), 2072-2098. https://doi.org/10.1177/1461444820925811
6. Lu, D., Zhang, D., Zhang, J., & Jin, G. (2023). Spatial-temporal transformer network for protecting person-of-interest from deepfaking.
7. Mehta, P., Jagatap, G., Gallagher, K., Timmerman, B., Deb, P., Garg, S., … & Dolan-Gavitt, B. (2023). Can deepfakes be created on a whim
8. Neethirajan, S. (2021). Beyond deepfake technology fear: on its positive uses for livestock farming.
9. Almars, A. M. (2021). Deepfakes detection techniques using deep learning: a survey. *Journal of Computer and Communications*, *9*(05), 20-35
10. Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, *9*(11).
11. Negi, S., Jayachandran, M., & Upadhyay, S. (2021). Deep fake: an understanding of fake images and videos. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *7*(3), 183-189.
12. Rebello, L., Tuscano, L., Shah, Y., Solomon, A., & Shrivastava, V. (2023), Detection of deepfake video using deep learning and MesoNet,In *2023 8th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1022-1026). IEEE.
13. Aduwala, S. A., Arigala, M., Desai, S., Quan, H. J., & Eirinaki, M. (2021). Deepfake detection using GAN discriminators. In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 69-77). IEEE.
14. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS).
15. Xia Z, Qiao T, Xu M, Wu X, Han L, Chen Y. Deepfake Video Detection Based on MesoNet with Preprocessing Module. Symmetry. 2022; 14(5):939.
16. Gupta, A., & Pandey, D. (2024). Unmasking the illusion: Deepfake detection through MesoNet. In *2022 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)* (pp. 1934-1938). IEEE.
17. Li, Z., Yang, W., Peng, S., & Liu, F. (2020). A survey of convolutional neural networks: Analysis, applications, and prospects.
18. Jogin, M., Mohana, M., Madhulika, M., Divya, G., Meghana, R., & Apoorva, S. (2018). Feature extraction using convolution neural networks (CNN) and deep learning. In *2018 3rd International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 2319-2323)
19. Hsu, C., Lee, C., & Zhuang, Y., (2018). Learning to detect fake face images in the wild. In *2018 IEEE International Symposium on Computer, Consumer and Control* (IS3C) (pp. 388-391).
20. Shi, W., Jiang, F., Liu, S., & Zhao, D. (2019). Image Compressed Sensing using Convolutional Neural Network. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 10.1109/TIP.2019.2928136. Advance online publication.
21. Huang, Y., Zhang, C., & Wang, J. (2022). GNN-EA: Graph neural network with evolutionary algorithm. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1476-1481). IEEE.
22. Pang L. (2022). Intelligent Big Information Retrieval of Smart Library Based on Graph Neural Network (GNN) Algorithm. *Computational intelligence and neuroscience*, *2022*, 1475069.
23. Ji, Y., Wang, J., Li, S., Li, Y., Lin, S., & Li, X. (2021). An anomaly event detection method based on GNN algorithm for multi-data sources. In *Proceedings of the 3rd ACM International Symposium on Blockchain and Secure Critical Infrastructure (BSCI '21)* (pp. 91–96). Association for Computing Machinery.
24. Hu, Z., Chen, K., & Xie, X. (2022). A graph neural network (gnn) algorithm for constructing the evolution process of rural settlement morphology. Security and Communication Networks, 2022, 1-10.
25. Aghasanli, A., Kangin, D., &Angelov, P., (2023) "Interpretable-through-prototypes deepfake detection for diffusion models,"*IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France, 2023, pp. 467-474.
26. Rana,M. S., &Sung, A. H.,(2020) "DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection,"*7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud),6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, New York, NY, USA, 2020, pp. 70-75
27. Vashishtha, S., Gaur, H., & Das, U. (2024). Optifake: Optical flow extraction for deepfake detection using ensemble learning technique. *Multimedia Tools and Applications*
28. Heidari, A., Navimipour, N. J., & Dag, H. (2024). A novel blockchain-based deepfake detection method using federated and deep learning models. *Cognitive Computation*.
29. Hubálovský, Š., Trojovský, P., Bacanin, N., & K, V. (2022). Evaluation of deepfake detection using YOLO with local binary pattern histogram. *PeerJ. Computer science*, *8*, e1086.
30. Agarwal, A., Agarwal, A., Sinha, S., Vatsa, M., & Singh, R. (2021). MD-CSDNetwork: Multi-domain cross stitched network for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1-8). IEEE
31. K, V., Trojovský, P., & Hubálovský, Š. (2023). VIOLA jones algorithm with capsule graph network for deepfake detection. *PeerJ. Computer science*, *9*, e1313.
32. Farid, H., & Bravo, M. J. (2010). Image Forensic Analyses that Elude the Human Visual System. *Proceedings of the SPIE, Human Vision and Electronic Imaging XV*, 7527, 75270S
33. Xu, G., Wu, Y., & Shi, Y. Q. (2016). Structural Design of Convolutional Neural Networks for Steganalysis. *IEEE Signal Processing Letters*, 23(5), 708-712.
34. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2017). FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv preprint arXiv:1803.09179*
35. Li, Y., & Lyu, S. (2019). Exposing DeepFake Videos By Detecting Face Warping Artifacts. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 46-52.
36. Qi, H., Su, H., Sun, B., & Wang, X. (2020). DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. *Proceedings of the 28th ACM International Conference on Multimedia*, 1711-1719
37. Jaleel, Q., Ali, I.H. (2023). MesoNet3: A Deepfakes Facial Video Detection Network Based on Object Behavior Analysis. In: Al-Bakry, A.M., *et al.* New Trends in Information and Communications Technology Applications. NTICT 2022. Communications in Computer and Information Science, vol 1764. Springer, Cham
38. Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114).