International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

DeepFake Image Detection

Mrs. Prajwal S. Gaikwad - Assistant Professor, Department of Computer Engineering
AISSMS Institute of Information Technology, Pune
prajwal.gaikwad@aissmsioit.org

Yadnesh Anil Patil
AISSMS Institute of Information Technology,
Pune Pune, Maharashtra, India
yadneshpatil572001@gmail.com

Ashutosh Manoj Pandey
AISSMS Institute of Information Technology,
Pune Pune, Maharashtra, India
ashutosh.pandey@aissmsioit.org

Naman Tarun Guntiwar
AISSMS Institute of Information Technology,
Pune Pune, Maharashtra, India
naman13235@gmail.com

Swaraj Ramdas Buchude
AISSMS Institute of Information Technology, Pune Pune,
Maharashtra, India
swarajrb07@gmail.com

***Abstract*—The growth of deepfakes in today's digital environment raises significant doubts regarding the genuineness and dependability of the content found. To overcome this new challenge, Developing an effective method in the context of detection of deep images. In this study, we conduct a comparative analysis of three varied convolutional neural networks (CNNs) for deepfake image detection. Our experimental results highlight the strengths and weaknesses of each CNN architecture.**

We deliberate on the consequences of our results in the context of deep image perception and show which models may be better for certain situations. We also address the challenges and limitations associated with deep learning, such as the arms race of deep learning technologies and tools. In conclusion, our work adds to the expanding body of knowledge regarding deep image detection by comparing three major CNN architectures. Our findings provide important guidance for researchers, prac- titioners, and policymakers working to improve the security and authenticity of content in an increasingly digital age. As deepfake technology continues to evolve, the information presented in this study sets the groundwork for development of more powerful and updated deepfake detection mechanisms.

***Keywords*—Deepfake, image detection, convolutional neural networks (CNNs), ResNet, InceptionV3, DenseNet, face forgery detection, GAN, forensics, deep learning, artificial intelligence, convolutional layers, pooling layers, fully connected layers, augmented data, accuracy, precision, recall, F1 measure, Area Under Receiver Operating Characteristic Curve.**

I. INTRODUCTION

Rapid advances in artificial intelligence and machine learn- ing have ushered in a new era of possibilities for digital media. But with these advances come new challenges, especially in the form of depth images. Deepfakes created using deep learning techniques can lead one to become overconfident in the other's interests, leading to potential deception and manipulation. With the advancement of deep learning technology, the need for effective research is now more important than ever. The rise of deepfakes has serious consequences, including misinformation, privacy threats, and security implications. Analysis of these completed images has evolved into a important endeavor to maintain the credibility involving digital content and maintain trust within visual media. Among the solutions, Convolutional Neural Networks (CNN) have emerged as an effective method due to their expertise in image analysis tasks. This research compares three different CNN architectures ResNet, InceptionV3, and DenseNet to explore the background of Deepfake image detection. Each design has unique characteristics that make it suitable for the task.

This research aims to understand the different strengths and limitations of CNN methods to solve this important problem by evaluating the performance of these networks in deep image analysis. This study provides an overview of the methodology and performance data, showing the implementation and training process of each CNN architecture. The subsequent presentation of evaluation results includes comparisons evaluated using metrics such as correctness, exactness, completeness, F1 score, and AUC-ROC, and other indicators. Together, these measurements provide a good understanding of each the model's capacity to differentiate between authentic images and depth pictures. The findings of this research will contribute to the ongoing fight against the development of radical images. This study aims to assist researchers, practitioners, and practitioners in their efforts to create effective and flexible in-depth mining by presenting the performance results of three popular CNN architectures. In the context of evolving deepfake technology, the insights gleaned from this comparative study potentially guide the advancement of advanced and robust solutions that increase the credibility and reliability of visual content in a growing Digital environment.

In addition to Deepfake nude photos, there are also nefarious or unlawful applications of Deepfakes, like disseminating false in- information, creating controversy, or many types of cybercrime. To deal with in the realm of Deepfake detection, tackling such threats has been excited by experts over recent years, leading to numerous instances of Deepfake detections. Some studies investigate the selected literature by focusing on the search process or performance analysis. However, further expansion of this area of research will help researchers and practitioners in the community to study and enable mutual

information collection of all types of deep fakes, including existing information that is not as good as before. research. To this finally, we display a qualitative literature review (SLR) of deep

learning in this article. Our aim is to identify and describe the similarities and differences between methods in current deep research practice. Our grants are summarized below. We analyze existing data in the field of deepfake. We report on the tools, methods and materials available for in-depth studies by conducting several research queries we present a classification that divides the deep search process into four groups and shows different new and original methods and their respective features. We assessed the depth of evidence from previous studies. We also use different metrics to assess the effectiveness of various deep learning methods. We share some observations and offer some suggestions for in-depth findings that may be useful for future research and practice in this field. The rest of this article follows this structure: Part II outlines the review section process by identifying questions of relevance. In Section III, we discuss the results of different studies in depth. Section 4 explains the details of overall analysis of the study, and Section 5 presents the challenges and limitations. Finally, Section 6 concludes the article.

II. LITREATURE REVIEW

[1] This paper presents a comprehensive analysis of the progress in deep learning technology pertaining to the production and identification of deepfake content. By categorizing scientific papers into methods for detecting fake images and facial videos, the survey delves into the utilization of meticulously crafted characteristics versus deep characteristics, temporal characteristics
across frames, and visual anomalies within video frames. It examines the difficulties, emerging patterns, and future directions in deepfake identifi- cation and multimedia forensics. The survey emphasizes the simplicity of generating deepfakes, the exceptional quality of manipulated videos, and the resulting influence on trust in media content. Moreover, the paper investigates the potential of generative adversarial networks (GANs) as a deep learning technique, for generating authentic deepfakes, as well as the development of detection strategies employing explain- able AI and white box models. This comprehensive analysis aims to improve comprehension and facilitate the progress of robust approaches to combat the proliferation of sophisticated deepfake media.

[2] The document investigates the utilization of Convolutional Neural Networks (CNN) in the domain of Forensics Face Detection. This involves the utilization of pre-trained weights from VGG-Face and the subsequent fine-tuning of these weights for the purpose of fake face classification. In order to facilitate the training and evaluation process, GANs such as PG-GAN and DC-GAN are employed to generate synthetic face images. To ensure a balanced dataset, we leverage data augmentation methods like flipping and rotation are implemented. The experiments conducted in this study make use of the CelebA dataset as well as data from the AI Chal- lenge contest. The efficacy of the proposed approach is assessed through AUROC analysis, which reveals a high level of accuracy. Among the different models tested, the VGG- Face VGG16 model consistently delivers the most optimal results. The paper highlights the importance of aligning the training and test datasets, as well as the value of leveraging deep learning networks for feature extraction and fine-tuning for classification. The contributions made by this research encompass the development of appropriate training datasets, the integration of deep learning networks, and the successful outcomes achieved in the AI Challenge contest. The primary focus of this study centers on GAN forensics face detection, specifically emphasizing the effectiveness of CNNs in identi- fying counterfeit faces and mitigating the risks associated with identity theft.

[3]The present paper advances a methodology for the identification of Deep Fakes, which pertain to counterfeit facial images or videos generated through the utilization of artificial intelligence. This technique involves scrutinizing incongruities in 3D head orientations. Deep Fakes are engendered by inserting artificially produced visages into original images, thus
giving rise to errors that can be detected by analyzing head positions. By comparing head orientations calculated using full set of facial landmarks with those derived solely from the critical facial region, discernible discrepancies in Deep Fakes are detected due to variations in landmark placements. These variations function as indicators in a classification procedure, wherein we utilize an SVM to establish a boundary for classifying authentic and manipulated images. The approach extracts features for classification by estimating head orientations and quantifying the disparities between orientations obtained from the central facial area and the entire face. Through experimentation, this investigation validates the effectiveness of this approach in the identification of Deep Fakes based on inconsistencies in head orientations, thereby providing a valuable tool for media forensics and the mitigation of the dissemination of deceptive content.

[4] The DeeperForensics-1.0 dataset constitutes a noteworthy novel approach to the field of face forgery recognition benchmarks. It consists of a substantial collection of 60,000 top-notch videos, thereby enhancing the accuracy of this field. Numerous benchmarks such as FaceForensics Benchmark and Celeb-DF have been introduced to detect face manipulation. The DF- VAE method improves the scalability and multimodality of face swapping techniques. The research delves into different variants and distortions in the training set to boost the accuracy of detecting face forgery. A user study compares the DF-VAE method with existing face manipulation techniques, showcas- ing promising outcomes in terms of realism and quality. The dataset's realism rating is remarkably high, with participants expressing their appreciation for its quality. DeeperForensics- 1.0 stands out due to its extensive scale, surpassing previous datasets and ensuring a diverse and demanding benchmark for detecting face forgery. The dataset encompasses meticu- lously collected source videos and fabricated videos generated through an end-to-end face swapping framework, validated by user studies for quality.

[5] The research introduces a new methodology for identifying fraudulent facial images or deepfakes through the utiliza- tion of a distinctive representation of images known as face X-ray. This particular representation emphasizes the merging boundaries within manipulated images while simultaneously concentrating on authentic

images, thus enabling the identifi- cation of facial alterations. A completely convolutional neural network was utilized to forecast face X-ray and categorize images as either authentic or blended. The proposed technique surpassed existing methods in detecting unfamiliar facial al- terations, thereby exemplifying noteworthy enhancements in accuracy. By equilibrating the losses of classification and face X-ray prediction with a weight of loss, the model accom- plished enhanced performance. The methodology managed to generalize the identification of various blending techniques, illustrating its effectiveness in discerning manipulated images. In general, the research presents a sturdy and original solution for detecting deepfakes, thereby providing a promising avenue for combating the dissemination of counterfeit images in digital media.

[1] The PDFexamines the complexities of of detecting deepfake videos, with a specific focus on the challenges presented by manipulated videos generated by artificial intel- ligence (AI). It highlights the critical requirement for effective algorithms for detection. The document investigates the two primary methods utilized in the creation of deepfakes: autoen- coders and GANs. It sheds light on their applications and the implications they have for strategies to detect such videos. The document also provides an outline of various indicators that can be employed to identify deepfake videos, including discrepancies in facial features, artifacts, head positioning, and eye blinking rates. Furthermore, the paper introduces a model that utilizes noise addition and blur for the purpose of detection. It showcases the performance of this model on the Celeb-DF dataset. By comparing this model with existing methods, the study eunderscores the critical role of continuous research and innovation in the field of deepfake detection, in order to efficiently counter the advancing sophistication of deepfake technologies. The research aims to enhance the efficiency of detection and develop new indicators to tackle the growing challenges presented by deepfake content. It underscores the importance of staying one step ahead of the malicious applications of AI-generated videos.

III. PROPOSED SYSTEM

Our proposed methodology for identifying instances of deepfake in images is depicted in Figure 1. In this in- vestigation, we have implemented two distinct classification techniques. Figure 1 demonstrate the application of the same technique and framework for processing the input data and also introduces an additional phase of analysis by means of a the classification step following post-processing is being questioned, which has undergone a integrated into the final output layer of the examined models. The aim of this secondary cycle of examination combined with the supplementary post-processing was intended to assess the influence of conducting principal component analysis on the efficiency of classifying deepfakes. For a more comprehensive understanding of the post-processing step, detailed explanations can be found in the concluding paragraphs of the evaluation subsection within this section.

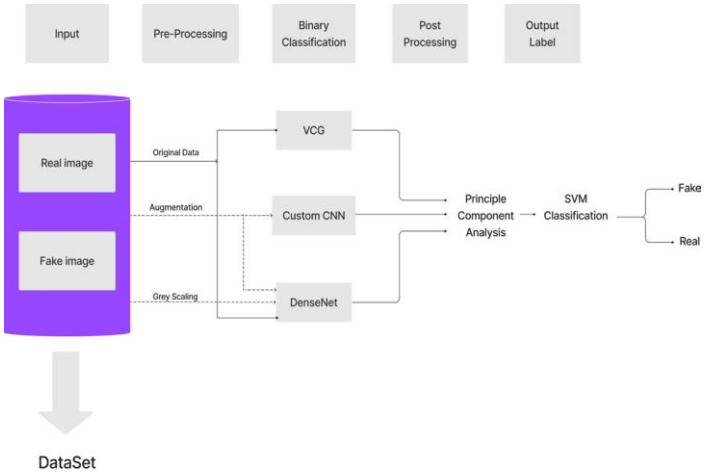

*Fig. 1. System Architecture*Due to the difficulties posed by the ever-increasing number of fake DeepFake images in forensic evaluation, we used 2 existing CNN frameworks, VGGFace and DenseNet, as well as a custom CNN, the main goal of which was to distinguish real data from fake data.

The provided data represents a dataset that has been sorted into two distinct groups: actual and fraudulent. These data have been enriched through the application of specific parameters for the purpose of training. The parameters utilized include a the rotation range for DenseNET is set to 20 degrees, while there is zero rotation was applied to the Custom CNN model. Additionally, a scaling factor of 1/255 was employed to reduce coefficients, and the shear range approximately 0.2 was randomly applied to introduce shearing transformations. Furthermore, A zoom range of 0.2 is utilized and was ran- domly implemented to enable zooming within images. Finally, randomized images were created through the use of flipping horizontally and vertically. After the augmentation process, the facial images underwent classification into two categories: fake or real. This classification was accomplished using three distinct models: Custom CNN, VGG, and DenseNET. For the purpose of our binary classification task, two classes were established. The first class, denoted as 0, represents the real category, encompassing normal, validation, and disguised facial images. The second class, denoted as 1, pertains to the fake category, which includes impersonator facial images.

The dataset known as the ”Real and Fake Face-Detection” was employed in training the trio of models with a learning rate set at 0.001 and a duration of 10 epochs. Subsequently, the accuracy on the test set reached determined by evaluating the testing dataset. To expand the dataset size, all initial images were flipped vertically and horizontally, resulting in a threefold expand in data.

A. VGGFace

VGG16 is a Convolutional Neural Network (CNN) that is one of the foremost computer vision models. It is also called

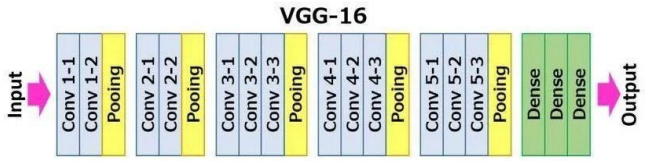

Fig. 2. VGG-16 Architecture

a ConvNet. A ConvNet is a multilayer artificial neural network designed for processing information, consisting of an input layer for receiving data, an output layer for generating results, and different hidden layers. The creators evaluated the networks and extended the depth using a very small architecture with 3 × 3 (3) convolution filters. The prior-art configurations showed a significant improvement. They heightened the complexity to 16 to 19 weight added layers, making it approximately — 138 Trainable Parameters. The VGG16 constitutes an object detection and classification (OD&C) algorithm. It has the capability to classify 1000 images from 1000 varied groupings achieving an accuracy rate of 92.7 percent. This is one of the most popular image classification algorithms and stands as easy to implement using transfer learning.
Here are the mathematical equations used in each component of the VGG16 model:

Here are the mathematical equations used in each component of the VGG16 model:

1. Convolutional Layer:
   Input: (X) (input feature map)
   Filter: (W_k) (filter weights for the (k)th filter)
   Bias: (b_k)(bias term for the (k)th filter) Output:
   $Y_k = ReLU(X * W_k + b_k)$

2. Max Pooling Layer:
   Input: (X) (input feature map)
   Pool size: ($P * P$) (pooling size)
   Stride: (S) (stride)
   Output: $Y_{\{i,jj,k\}} = max_{\{m,n\}\{X}$ $\{(i * S + m),(jj * S + n),k\}$
3. Fully Connected Layer:
   Input: (X) (input vector)
   Weight matrix: (W) (weights) Bias:
   (b) (bias term)
   Output: $Y = \{ReLU\}(XW + b)$

4. Softmax Activation:
   Input: (Z) (raw scores before softmax)
   Output: $P_i = \frac{e^{Z_i}}{\sum_{jj} e^{Z_{jj}}}$ (probability for class i)
5. Flattening:
   Input: 3D tensor from the last pooling layer Output:
   Flatten the 3D tensor into a 1D vector

These equations represent the operations performed in each layer of the VGG16 model, including convolution, pooling, fully connected layers, softmax activation, and flattening.

sequentially in traditional networks, DenseNet links each layer towards other in a "feed-forward" manner. This dense connectivity enhances feature propagation, reduces parameter count, and improves model performance. DenseNet is consists of dense blocks that contain convolutional stratum and transition stratum that minimize channel count. The network terminates at average pooling across all spatial locations, where a fully connected layer is connected for purpose of classification. In summary, DenseNet accomplishes excellent performance in tasks related to classifying images by leveraging feature reuse and increasing each component. information flow. Here are the mathematical equations used in

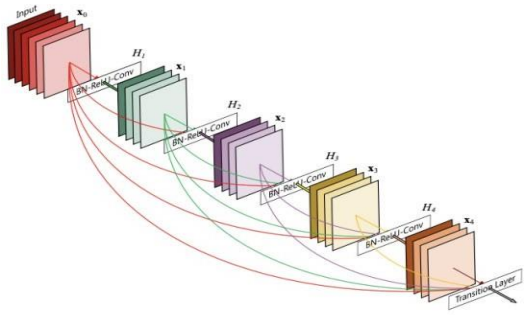

Fig. 3 . Dense Convolutional Network (DenseNet)

DenseNet model:

1. Dense Block:
Input: X_l (input feature maps to the l th dense block) Combination of all preceding maps: $[X_0, X_1, \ldots, X_{l-1}]$ Output: H_l = ReLU (BatchNorm(Conv($[X_0, X_1, \ldots, X_{l-1}]$)))

2. Transition Layer:
Input: X_l (input feature maps from the dense block) Ouput feature maps: X_{l+1} = ReLU (BatchNorm(Conv(X_l))) Compression factor: Employed for diminishing the quantity of feature maps by a factor of $\theta\theta$.

3. Global Average Pooling:
Input: Feature maps from the last dense block
Output: Spatial dimensions are reduced to 1 × 1, and the channel dimension remains unchanged.

4.Fully Connected Layer:

B. DenseNet

DenseNet's deep learning architecture focuses on layer- tolayer connectivity. Instead of connecting layers Input: Flattened feature maps from the global average pooling layer Output: Y =

Softmax($XXWW + b$) (final classification probabilities)

In DenseNet, each layer receives feature maps from all preceding layers transmit their own feature maps to all subsequent layers, fostering improved feature reuse and gradient flow within this dense connectivity pattern, leading to improved performance compared to traditional architectures.

V. CONCLUSION

IV. RESULT

Model	Neural Network		
Metrices	Accuracy	Precision	Recall
Custom Model	0.931	0.89	0.89
Custom Model with Augmented Data	0.907	0.84	0.84
VGGFace	0.96	0.95	0.95
DenseNet	0.956	0.94	0.93
DenseNet with Augmented Data	0.871	0.79	0.73
DenseNet with Gray Scale Images	0.956	0.95	0.94

TABLE I. Neural Network results

Model	SVM after PCA		
Metrices	Accuracy	Precision	Recall
Custom Model	0.974	0.972	0.976
Custom Model with Augmented Data	0.91	0.906	0.916
VGGFace	0.995	0.995	0.995
DenseNet	0.984	0.987	0.981
DenseNet with Augmented Data	0.863	0.861	0.864
DenseNet with Gray Scale Images	0.503	0.50	0.516

TABLE II. SVM after PCA result

The VGGFace architecture, achieved the highest level of accuracy at 96%. Nonetheless, this particular model demands a significant amount of computational resources and necessitates sophisticated processors to facilitate training on augmented data. In contrast, the DenseNet architecture yielded comparable outcomes, as evidenced in Table 1, while also exhibiting greater efficiency and aligning with the assertions made.The introduction of data augmentation led to a decrease in accuracy for the DenseNet model, approximately 8-10%, which is expected since data augmentation intensifies the difficulty of training samples.The possibility of achieving even superior outcomes exists through the model's training phase for a greater number of epochs. The performance of the DenseNet architecture on grayscale images demonstrates that color does not influence the model's capacity to classify GAN-generated images. The Custom Model also produced respectable results, with its computational efficiency lying between that of VGGFace and DenseNet. When compared to DenseNet, the Custom Model exhibited superior results on augmented data. The ROC plot in Fig. 1 provides a visual representation of the relative performance of all models.

DeepFakes are here to stay and in doing so have changed our perception of reality forever. An immense challenge in developing forensic methods to detect real versus fake images and videos is that once papers are published on new innovative approaches or methods are shared via open access, these flaws are immediately incorporated in subsequent iteration pertaining to DeepFake generation techniques. Even with models with accuracy as high as 97% are not enough. Similar to the medical domain, it is the ones that are missed that represent the larger problem - i.e., 3% of billions of images on Google or Facebook platforms would represent an immense loss of trust from users of these interfaces. Our results show that state-of-art CNNs are now able to distinguish with minimal mis- classification inaccuracies between fake and real data.

However, the identifying these addressing minimal inaccuracies remains a pivotal focus of research. Recent endeavors have concentrated on enhancing DeepFake generation algorithms by incorporating intricately crafted noise into digital images or videos. This noise, imperceptible to the human eye, aims to deceive face detection algorithms effectively. Ultimately, this is a battle now between human ingenuity and the ubiquitous pervasive presence of machines which have qualities which allow them to become iteratively intelligent. Future work would iIncorporating unsupervised clustering techniques like auto- encoders enables the exploration of whether genuine and counterfeit images segregate distinctly. Additionally, these methods enhance the clarity and comprehensibility of our models through the utilization of CNN visualization techniques.

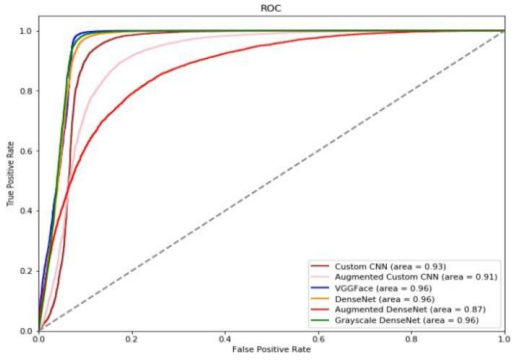

Fig 4. ROC Curve

VI. REFERENCES

[1] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. Computer Vision and Image Understanding. https://doi.org/10.1016/j.cviu.2022.103525

[2] Do, Nhu-Tai & Na, In & Yang, Hyung-Jeong & Lee, Guee-Sang & Kim, S.H.. (2018). Forensics Face Detection From GANs Using Convolutional Neural Network. https://www.researchgate.net/publication/327905310

[3] Yang, X., Li, Y., & Lyu, S. (2018). Exposing Deep Fakes Using Inconsistent Head Poses. arXiv preprint arXiv:1811.00661. https://doi.org/10.48550/arXiv.1811.00661

[4] Jiang, L., Li, R., Wu, W., & Qian, C. (2020). DeeperForensics-1.0: A Large-Scale Dataset for Real- World Face Forgery Detection. Nanyang Technological University SenseTime Research. arXiv preprint arXiv:2001.03024. https://doi.org/10.48550/arXiv.2001.03024

[5] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2019). Face X-ray for More General Face Forgery Detection. arXiv:1912.13458. https://doi.org/10.48550/arXiv.1912.13458

[6] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov and A. S. Smirnov, "Methods of Deepfake Detection Based on Machine Learning," *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, St. Petersburg and Moscow, Russia,2020,pp.408-411,doi:10.1109/EIConRus49466.2020.9039057.

[7] Coccomini, D., Messina, N., Gennaro, C., & Falchi, F. (2022). Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In International Conference on Multimedia Modeling (pp. 267-278). Springer, Cham. https://doi.org/10.1007/978-3-031-06433-3_19

[8] Kumar, A., & Bhavsar, A. (2020). Detecting Deepfakes with Metric Learning. arXiv preprint arXiv:2003.08645. https://doi.org/10.48550/arXiv.2003.08645

[9] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2020). Video Face Manipulation Detection Through Ensemble of CNNs. arXiv preprint arXiv:2004.07676. https://doi.org/10.48550/arXiv.2004.07676

[10] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. In 10th IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE. https://doi.org/10.1109/WIFS.2018.8630761

[11] A. A. Pokroy and A. D. Egorov, "EfficientNets for DeepFake Detection: Comparison of Pretrained Models," *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, St. Petersburg, Moscow, Russia, 2021, pp. 598- 600, doi: 10.1109/ElConRus51938.2021.9396092.

[12] 78. https://doi.org/10.4236/jcc.2021.99009

[13] Guo, Y., Jiao, L., Wang, S., Wang, S., & Liu, F. (2018). Fuzzy Sparse Autoencoder Framework for Single Image per Person Face Recognition. IEEE Transactions on Cybernetics,48(6),1863-1876. https://doi.org/10.1109/TCYB.2017.2739338

[14] Goel, R., Mehmood, I., & Ugai, H. (2021). A Study of Deep Learning- Based Face Recognition Models for Sibling Identification. Sensors, 21(15), 5068. https://doi.org/10.3390/s21155068