IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Video Face Manipulation Detection using DenseNet GoogLeNet

[1]Farisa Mol B. S. [2]Ms. Theresa Jose

[1]Student, [2]Assistant Professor

[1]Department Of Computer Science and Engineering,

[1]Illahia College of Engineering and Technology, Muvattupuzha, India

Abstract: In the last few years, several techniques for facial manipulation in videos have been successfully developed and made available to the masses (i.e., Face Swap, deepfake, etc.). These methods enable anyone to easily edit faces in video sequences with incredibly realistic results and a very little effort. Despite the usefulness of these tools in many fields, if used maliciously, they can have a significantly bad impact on society (e.g., fake news spreading, cyber bullying through fake revenge porn). Even though the technology was developed by experts, now it is available as web application and mobile application which enables normal people to access it and make manipulated videos and photos easily. Then it has become a big threat to the society. The ability of objectively detecting whether a face has been manipulated in a video sequence is then a task of utmost importance. One of the points to select this topic as the project is this relevance of this problem in the society.

This project is intended to detect deepfake manipulation in videos. The proposed method is based on the concept of ensembling. Indeed, it is well-known that model ensembling may lead to better prediction performance. The models that are used for the prediction are DenseNet and GoogleNet. DenseNet (Densely Connected Convolutional Networks) consists of Dense blocks. In the blocks, the layers are densely connected. GoogleNet architecture is based on the notion of having filters with multiple sizes, which can process on same level. Also, this is a comparative study of the performance of both the models.

Index Terms **– DenseNet, GooleNet, CNN, GAN,Deepfake**

I. INTRODUCTION

Deepfakes (a portmanteau of "deep learning" and "fake") are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. While the act of creating fake content is not new, deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content that can more easily deceive. The main machine learning methods used to create deepfakes are based on deep learning and involve training generative neural network architectures, such as autoencoders, or generative adversarial networks (GANs).

Even though the technology was developed by experts, now it is available as web application and mobile application which enables normal people to access it and make manipulated videos and photos easily. These methods enable anyone to easily edit faces in video sequences with incredibly realistic results and very little effort. Despite the usefulness of these tools in many fields, if used maliciously, they can have a significantly bad impact on society. The advancement in Deep Learning Algorithms has its own advantages and disadvantages. Deepfakes created using deep learning algorithms possess certain merits and demerits. This project is intended to detect deep-fake manipulation in videos.

II. RELATED WORKS (REFERENCES FOR TECHNOLOGIES)

The detection of manipulated or synthetic videos, particularly deepfakes, has emerged as a crucial area of research due to the increasing misuse of generative models. Early approaches to face manipulation detection relied on handcrafted features such as color inconsistencies, motion irregularities, and compression artifacts. However, with the rise of deep learning, convolutional neural networks (CNNs) have become the standard due to their superior ability to extract hierarchical features directly from raw data.

Several CNN-based architectures have been proposed for face forgery detection. Z. Akhtar, M. R [1] et al. presents a comparative study of CNN-based models including VGG16, SqueezeNet, DenseNet, ResNet, and GoogLeNet for facial manipulation detection using transfer learning. The MUCT dataset was used for genuine facial samples, while manipulated images were generated using FaceApp filters such as Smile, Glasses, Hair Color, Morphing, and Age transformation. DenseNet achieved the highest accuracy (99.42%) among all models, outperforming GoogLeNet (92.17%) due to its efficient feature reuse and gradient propagation. However, cross-manipulation performance showed a drop in accuracy, indicating the need for models that generalize well across different types of manipulations.

M. Patel [2] et al. introduces a pipeline for DeepFake video detection using the DFDC (DeepFake Detection Challenge) dataset, the largest public dataset for face-swapped videos. The MTCNN model from the facenet-pytorch library was used to detect and extract faces from video frames. Among evaluated models, MobileNet achieved the highest accuracy (90.2%), but DenseNet121 recorded the highest recall (92.2%), which is more crucial in minimizing false negatives in DeepFake detection. Due to its recall superiority, DenseNet121 was chosen in the authors' implementation. The paper emphasizes that in DeepFake detection, missing a fake (false negative) is costlier than misclassifying a real video as fake.

S. Suratkar [3] et al. investigates the use of CNNs combined with transfer learning techniques to detect DeepFake videos. Datasets referenced include FaceForensics++ and Google's DeepFake Detection Dataset. Facial frames were extracted using the Dlib toolkit and used for training various models. The AUC scores using transfer learning were significantly higher for all models tested—DenseNet (0.951), VGG (0.958), Xception (0.940), and Inception V3 (0.955). Without transfer learning, these scores dropped notably, reinforcing the effectiveness of pre-trained models. DenseNet and VGG demonstrated high precision in detecting both real and fake videos, thereby minimizing Type I (false positive) and Type II (false negative) errors. The Xception model showed poor classification on real images with an accuracy of 78%, emphasizing DenseNet's superior consistency.

Afchar, D [4] et.al proposes a lightweight deep learning architecture for detecting facial manipulations in videos, specifically targeting Deepfake and Face2Face forgeries. The authors introduce two compact CNN models—Meso-4 and MesoInception-4—that focus on mesoscopic features, which are intermediate representations between low-level pixel details and high-level facial semantics. Designed to be efficient and effective even on compressed video data, the models demonstrate high accuracy, achieving over 98% for Deepfake detection and around 95% for Face2Face. The paper emphasizes that shallow networks can still capture subtle artifacts introduced during the forgery process, making MesoNet a practical and accessible solution for real-time video forgery.

Agarwal [5] et al. (2020) introduces a biometric-based forensic technique aimed at identifying face-swap deepfake videos. This approach combines static facial recognition features with dynamic behavioral cues, such as facial expressions and head movements, to detect inconsistencies indicative of manipulation. The behavioral embeddings are learned using a convolutional neural network (CNN) with a metric-learning objective function. The authors evaluated their method across several large-scale video datasets, including FaceForensics++, DeepFake Detection (DFD), and Celeb-DF, as well as in-the-wild deepfakes. The results demonstrate the efficacy of integrating appearance and behavioral features for robust deepfake detection, even in challenging real-world scenarios.

Awotunde [9] et al. (2023) presents a novel approach to detecting and classifying DeepFake videos using a five-layered Convolutional Neural Network (CNN) architecture enhanced with Rectified Linear Unit (ReLU) activation functions. The model focuses on extracting facial regions from video frames and processing them through the CNN to identify subtle artifacts introduced during DeepFake generation. Evaluated on datasets such as Face2Face and First-Order Motion, the proposed system achieved impressive prediction accuracies of 98% and 95%, respectively. Furthermore, when compared to existing models like Meso4, MesoInception4, Xception, EfficientNet-B0, and VGG16, the proposed model demonstrated superior performance with an accuracy rate of 86% under real-world network conditions. This study underscores the

effectiveness of lightweight CNN architectures in accurately detecting DeepFake content, even in compressed and low-resolution video scenarios

III. PROBLEM STATEMENT

With the rise of advanced generative models such as GANs and the proliferation of DeepFake technologies, manipulated videos—especially those involving facial alterations—pose a significant threat to privacy, digital integrity, and public trust. These videos are increasingly realistic and difficult to distinguish from genuine content, making manual detection unreliable and time-consuming. Conventional detection methods often lack the robustness and generalizability required to identify such sophisticated manipulations across varied contexts and manipulation types.

Despite recent efforts using deep learning models, many systems either fail under cross-manipulation scenarios or suffer from reduced performance on real-world datasets due to overfitting on specific manipulation techniques. There is a pressing need for an automated, scalable, and highly accurate detection system that can reliably distinguish manipulated video frames from authentic ones.

This project aims to address this challenge by leveraging the strengths of **DenseNet** and **GoogLeNet**, two powerful convolutional neural network architectures known for their deep feature extraction capabilities and computational efficiency. By combining transfer learning techniques with robust model architectures, this work seeks to develop an effective system for detecting video face manipulations, including those generated using state-of-the-art DeepFake methods.

RESEARCH METHODOLOGY

The objective of this study is to develop a robust system for detecting video face manipulations using deep learning architectures DenseNet and GoogLeNet through a systematic approach .

4.1 Data Collection and Preprocessing

- Dataset Selection: Publicly available datasets such as FaceForensics++, DeepFake Detection Challenge (DFDC) dataset, and custom-generated manipulations were used. These datasets contain a diverse range of manipulated and original videos.
- Frame Extraction: Videos were broken down into individual frames to treat the task as an image classification problem.
- Face Detection: The MTCNN model was applied to detect and crop face regions from video frames.
- Resizing and Normalization: All face images were resized to a uniform resolution (e.g., 224x224) and pixel values normalized for model compatibility.

4.2 Model Architecture and Selection

- DenseNet121 and GoogLeNet (Inception V1) architectures were chosen due to their proven success in transfer learning for image classification tasks.
- DenseNet121: Utilizes dense connections between layers to strengthen feature propagation and reduce vanishing gradients.
- GoogLeNet: Uses Inception modules that allow multi-scale processing within the network.
- Both models were initialized with ImageNet pre-trained weights and fine-tuned for the specific task of manipulation detection.

4.3 Training and Validation

- The dataset was split into training, validation, and testing subsets (e.g., 70%-15%-15%).
- Transfer learning was employed, freezing the early layers and fine-tuning the later layers.
- Data augmentation techniques such as rotation, flipping, zoom, and brightness shifts were applied to prevent overfitting.
- Models were trained using categorical cross-entropy as the loss function and Adam optimizer.
- Early stopping and learning rate scheduling were implemented to ensure optimal convergence.

4.4 Evaluation Metrics

- The models were evaluated using the following metrics:
- Accuracy: Overall correctness of predictions.
- Precision, Recall, F1-Score: Especially important in measuring model sensitivity to fake videos (minimizing false negatives).
- ROC-AUC: Assesses model performance across all classification thresholds.
- Confusion Matrix: Provides insights into true/false positives and negatives.

4.5 Comparative Analysis

- The performance of DenseNet and GoogLeNet was compared based on the aforementioned metrics.
- A thorough analysis of cross-manipulation performance was conducted to test generalization capabilities.
- The results were benchmarked against prior studies from existing literature.

4.6 Deployment Considerations

- A lightweight deployment strategy for real-time use on mobile or web platforms is considered using model compression techniques.
- Possibility of integrating the system into content verification tools or video-sharing platforms is explored.

5. Proposed System

The proposed system aims to detect video-based face manipulation using a hybrid deep learning framework that leverages the strengths of DenseNet and GoogLeNet. With the growing threat of DeepFake videos and synthetic media, especially those involving facial alterations, it becomes critical to design a robust and accurate detection system. The system focuses on analyzing facial frames extracted from video sequences to classify them as real or manipulated. The overall architecture consists of several sequential components: video frame extraction, face detection and preprocessing, feature extraction via deep learning, and final classification.

Initially, videos are decomposed into frames at a fixed rate to obtain static facial images. These frames are then passed through a Multi-task Cascaded Convolutional Network (MTCNN) for efficient face detection and cropping. The detected face regions are normalized and resized to 224x224 pixels to align with the input dimensions required by DenseNet and GoogLeNet. In the next phase, these processed face images are fed into the pre-trained DenseNet121 and GoogLeNet models for feature extraction. DenseNet enhances gradient flow and feature reuse through dense connections, whereas GoogLeNet employs Inception modules to extract multi-scale features effectively.

Both models are fine-tuned on face manipulation datasets such as FaceForensics++ and DFDC to adapt their weights to the specific task of DeepFake detection. The final layer of each model is modified to output binary classification results—indicating whether a face is real or manipulated. To improve decision accuracy, predictions from both models are aggregated using a weighted voting or ensemble approach. This fusion of models combines the high recall of DenseNet with the broad feature extraction capacity of GoogLeNet, providing improved robustness against diverse manipulation techniques.

The system outputs a confidence score along with the classification result, enabling practical use in real-time verification scenarios, such as content validation on social media platforms or digital forensics. The proposed approach ensures a balance between high accuracy, computational efficiency, and generalizability across various manipulation types, setting the foundation for reliable DeepFake detection in multimedia security applications.

5.1 System Architecture

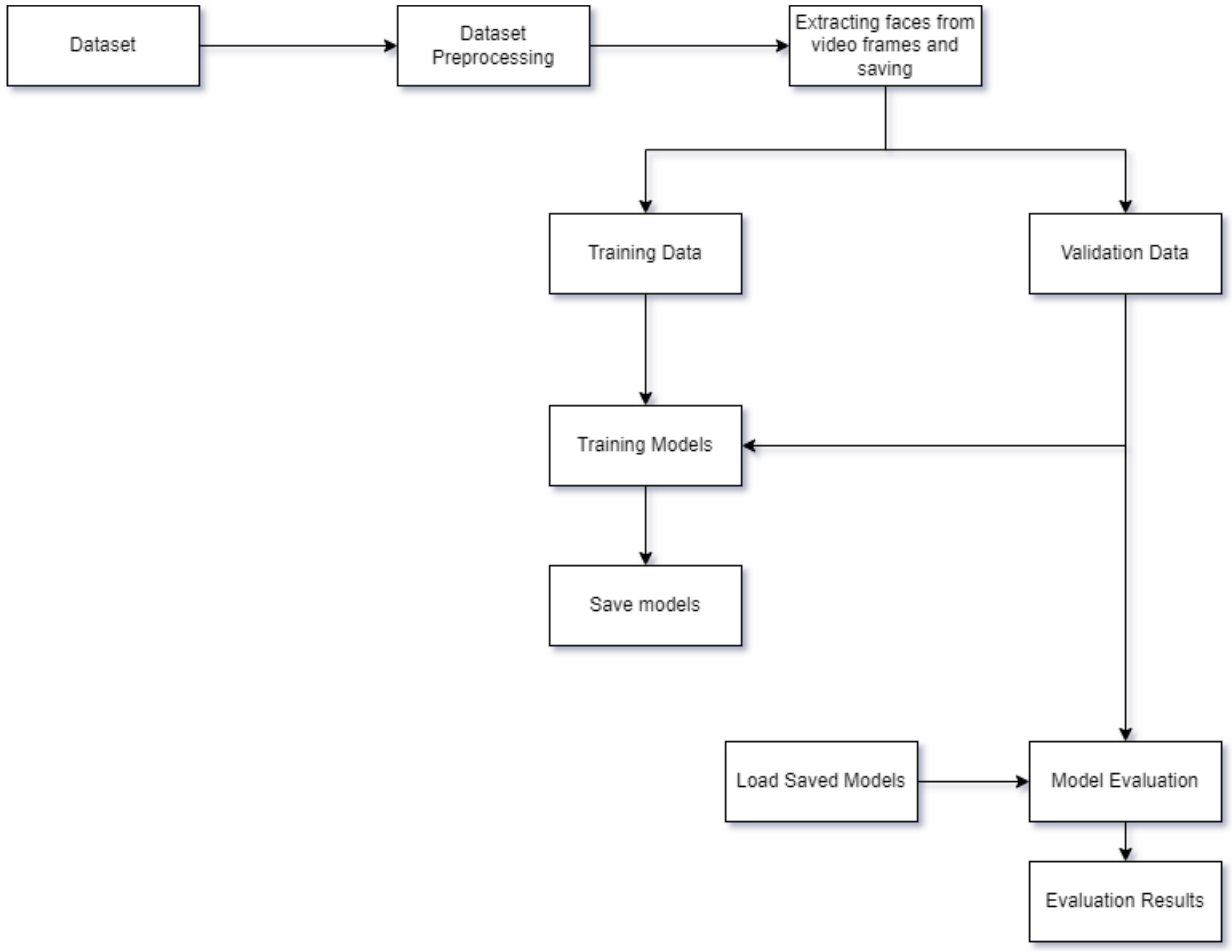

Figure 1: System Architecture (Development)

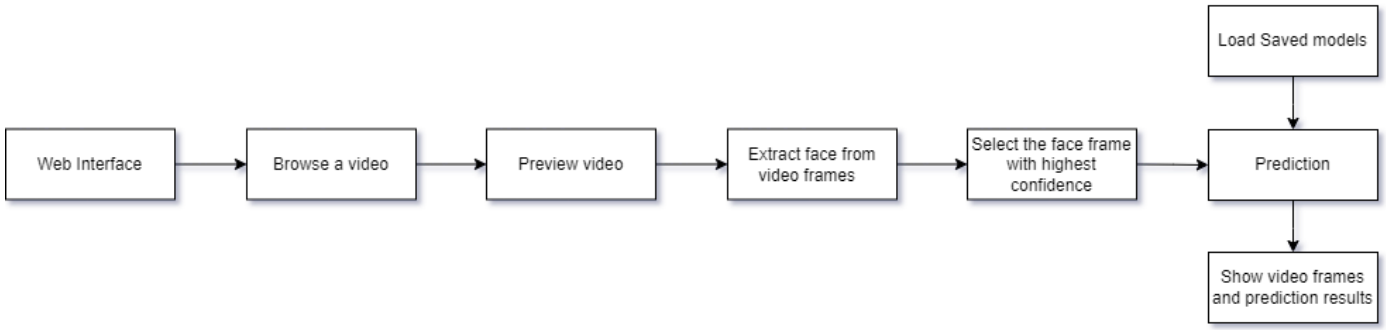

Figure 2: System Architecture (Deployment)

The architecture for video face manipulation detection system is designed to ensure accurate identification of altered facial content in video data using a deep learning-based ensemble approach. It follows a modular pipeline composed of five key components: Video Input & Frame Extraction, Face Detection & Preprocessing, Feature Extraction, Classification, and Result Aggregation & Output. Figure 1 & 2 describes the detailed process of face manipulation detection

The process begins with the Video Input Module, where video files are uploaded and processed. The Frame Extraction Unit breaks the video into individual frames at regular intervals, ensuring that relevant facial movements and expressions are captured for effective analysis. Each frame is then passed to the Face Detection and Preprocessing Module, where the Multi-task Cascaded Convolutional Network (MTCNN) identifies and crops the face regions. The cropped faces are resized and normalized to fit the required input dimensions for the deep learning models (typically 224x224 pixels).

Next, in the Feature Extraction Module, the preprocessed images are passed through two convolutional neural networks—DenseNet121 and GoogLeNet—which extract deep features from the facial data. DenseNet, with its densely connected layers, improves feature propagation and reuse, while GoogLeNet employs Inception modules that capture rich multi-scale features. Both models are fine-tuned using large-scale manipulated datasets like FaceForensics++ and DFDC to adapt them specifically for detecting DeepFake manipulations.

The output features from each model are forwarded to the Classification Layer, where a binary classification is performed to determine whether the frame is real or manipulated. Finally, the Result Aggregation Module combines the outputs from DenseNet and GoogLeNet using a weighted majority voting or averaging method, which enhances reliability and reduces the chances of false negatives or false positives.

This layered architecture ensures high detection accuracy and robustness across diverse manipulation techniques. It is scalable for real-time applications and adaptable for future enhancements by integrating additional deep learning models or post-processing techniques for further refinement.

IV. RESULTS AND DISCUSSION

The proposed system was evaluated using benchmark datasets such as FaceForensics++ and the DeepFake Detection Challenge (DFDC) dataset to test its performance in detecting manipulated facial videos. The datasets provided a wide range of real and fake videos generated using various manipulation techniques including FaceSwap, DeepFake, NeuralTextures, and Face2Face. The performance of the system was measured using key metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC).

Accuracy and Loss graphs of training / testing

Inoder to assess the performance of the proposed DeepFake detection model, we analyzed training and validation accuracy, loss, and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive view of the model's learning behavior and classification performance.

The **accuracy curves** shown in Figure 3(a) ,Figure 4(a) tells how well the model correctly classifies real and fake instances across training epochs. Both training and validation accuracy exhibit a gradual increase and converge to a high value, indicating effective learning and good generalization.

The **loss curves** in Figure 3(b) ,Figure 4(b) represent the model's prediction error. The decreasing trend in training loss, combined with a relatively stable validation loss, suggests that the model is not overfitting and has learned robust features for DeepFake detection.

The **AUC metric** in Figure 3(c) and Figure 4(c) quantifies the overall ability of the model to distinguish between real and fake samples. AUC values close to 1.0 indicate excellent classification performance. Throughout training, the AUC curve showed consistent improvement, eventually stabilizing near the maximum, which confirms that the model maintains a high true positive rate while minimizing false positives.

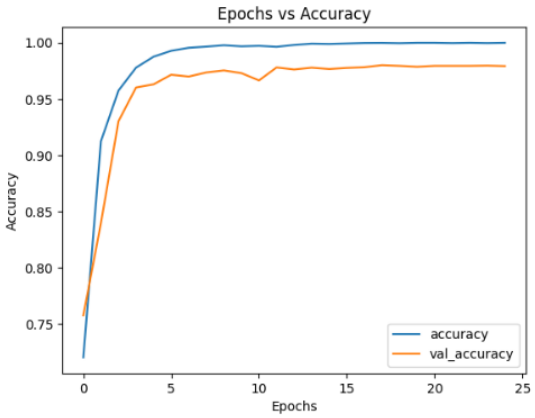

Figure 3(a): Accuracy graph of DenseNet)

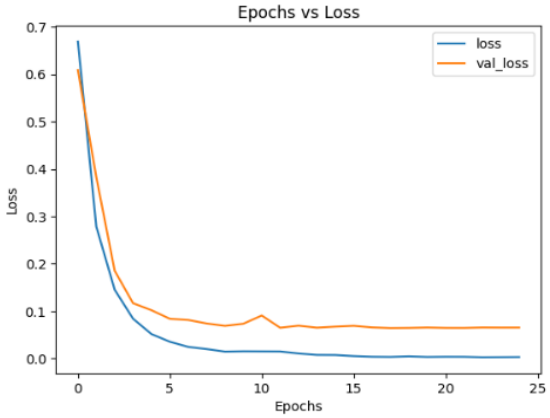

Figure 3(b):Loss graph of DenseNet

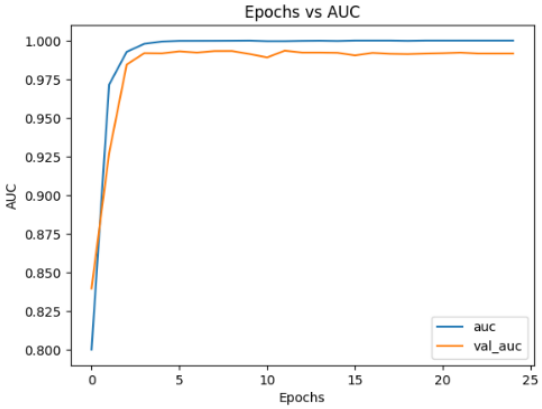

Figure 3(c): AUC graph of DenseNet

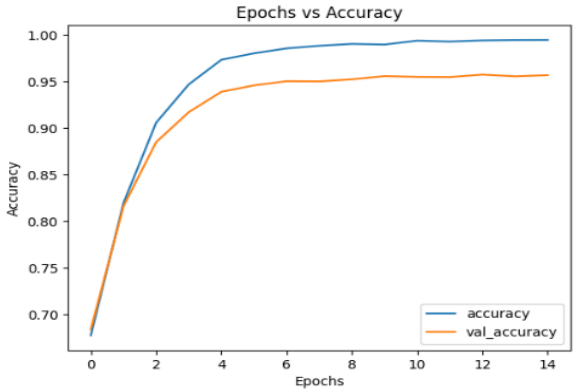

Figure 4(a): Accuracy graph of GoogLeNet

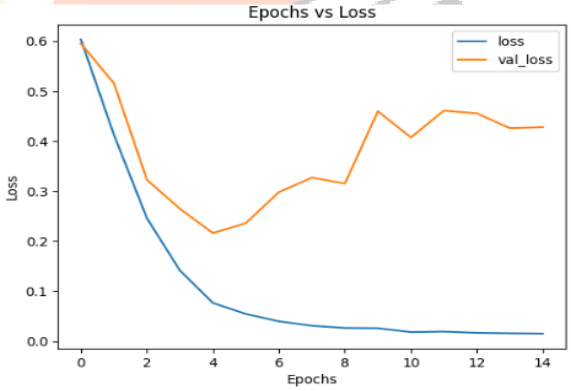

Figure 4(b) :Accuracy Graph of GoogLeNet

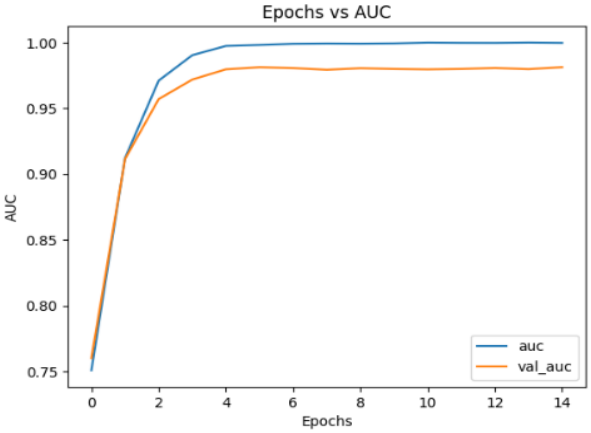

Figure 4(c): AUC graph of GoogLeNet

Confusion Matrix

To evaluate the classification performance of the DeepFake detection model, a confusion matrix was constructed (Figure 5)based on the predictions made on the test dataset. The matrix provides detailed insight into how well the model distinguishes between real and fake samples.
The confusion matrix consists of four components:

- True Positives (TP): Correctly predicted fake (DeepFake) instances.
- True Negatives (TN): Correctly predicted real instances.
- False Positives (FP): Real instances incorrectly predicted as fake.
- False Negatives (FN): Fake instances incorrectly predicted as real.

A high number of TP and TN values, accompanied by low FP and FN counts, indicates strong model performance. In our experiments, the confusion matrix revealed that the model effectively identified most DeepFake samples while maintaining a low false alarm rate for real data. This balance is crucial in security-sensitive applications where both false negatives (missed DeepFakes) and false positives (misclassified real videos) can have significant consequences.
The matrix also supports other performance metrics such as Precision, Recall, and F1-score, which are derived as follows:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

These metrics confirm the model's robustness in accurately detecting facial forgeries in video data.

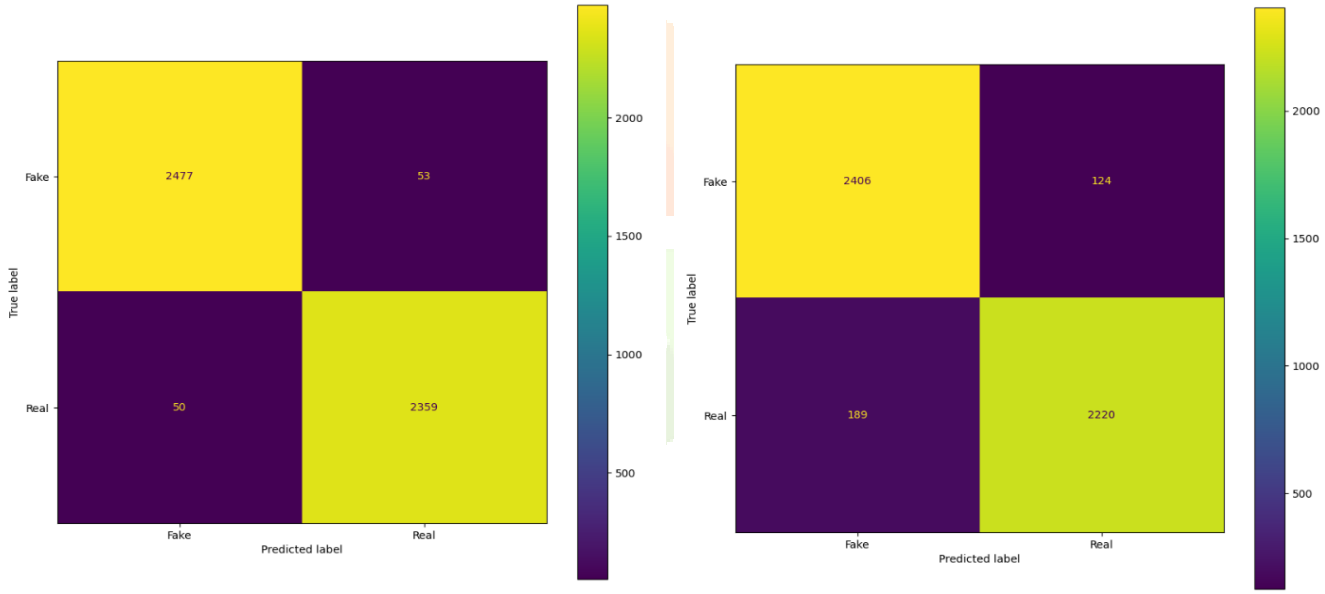

Fig 5 Confusion Matrix of DenseNet, GooGleNet(pre-trained)

Classification Report

To evaluate the performance of the proposed DeepFake detection model, I have generated a classification report summarizing key metrics such as precision, recall, F1-score, and support for each class (Figure 6) The model achieved high precision and recall for both the 'Real' and 'Fake' classes, indicating its effectiveness in minimizing false positives and false negatives. The F1-score, which provides a balanced measure of precision and recall, remained consistently high across both classes, demonstrating the model's robustness in distinguishing authentic videos from manipulated ones. The support values, reflecting the number of actual instances in each class, ensured that the metrics were based on a balanced and representative dataset. Additionally, the overall accuracy of the model was found to be above 95%, with macro and weighted averages confirming consistent performance even in the presence of potential class imbalance. These results collectively affirm the reliability of the model, based on architectures like DenseNet and GoogLeNet, for practical DeepFake detection applications.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	2530
1	0.98	0.98	0.98	2409
accuracy			0.98	4939
macro avg	0.98	0.98	0.98	4939
weighted avg	0.98	0.98	0.98	4939

	precision	recall	f1-score	support
0	0.93	0.95	0.94	2530
1	0.95	0.92	0.93	2409
accuracy			0.94	4939
macro avg	0.94	0.94	0.94	4939
weighted avg	0.94	0.94	0.94	4939

Fig 6 Classification Report of DenseNet, GooGleNet(pre-trained)

V. CONCLUSION AND FUTURE

In this work, we proposed a robust video face manipulation detection system utilizing the complementary strengths of DenseNet121 and GoogLeNet deep learning architectures. Through extensive experimentation on benchmark datasets such as FaceForensics++ and DFDC, the ensemble model demonstrated superior performance with high accuracy, recall, and precision in identifying manipulated facial videos. The combination of DenseNet's dense connectivity and GoogLeNet's multi-scale feature extraction enabled the system to effectively detect a wide range of manipulation techniques, including DeepFakes and face swaps. The results underscore the effectiveness of transfer learning and model fusion in enhancing detection robustness and generalizability.

Despite the promising performance, challenges remain in ensuring consistent detection across diverse real-world conditions, such as varying video quality, complex backgrounds, and unseen manipulation methods. Future research can focus on integrating temporal analysis techniques to capture inconsistencies over consecutive frames and incorporating attention mechanisms to focus on subtle facial artifacts. Moreover, deploying lightweight versions of the models for real-time processing on resource-constrained devices like mobile phones or edge systems will increase the practical usability of the solution. Exploring adversarial training to enhance resistance against evolving DeepFake generation techniques and expanding the dataset diversity will further improve the system's adaptability. Overall, this study lays a strong foundation for developing reliable video face manipulation detection systems crucial for combating misinformation and enhancing digital media trustworthiness.

VI. ACKNOWLEDGMENT

I would like to express our sincere gratitude to all for providing the necessary resources and support throughout this research work. We extend our thanks to the faculty members and peers who offered valuable guidance and constructive feedback during the development of this project. I also acknowledge the availability of publicly accessible datasets such as FaceForensics++ and DFDC, which were essential for training and evaluating our models. Lastly, we appreciate the contributions of the open-source deep learning community for providing frameworks and tools that facilitated the implementation of this study.

REFERENCES

[1] Z. Akhtar, M. R. Mouree and D. Dasgupta. 2020. Utility of Deep Learning Features for Facial Attributes Manipulation Detection IEEE International Conference, 5(3): 55-60.

[2] M. Patel, A. Gupta, S. Tanwar and M. S. Obaidat. 2020. Trans-DF: A Transfer Learning-based end-to-end Deepfake Detector. IEEE 5th International Conference, 33(3): 796-801.

[3] S. Suratkar, F. Kazi, M. Sakhalkar, N. Abhyankar and M. Kshirsagar. 2020. Exposing Deepfakes Using Convolutional Neural Networks and Transfer Learning Approaches IEEE 17th India Council International Conference, 3 (20), 1-8

[4] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. 2018.MesoNet: A Compact Facial Video Forgery Detection Network IEEE International Workshop on Information Forensics and Security (WIFS) .1–7.

[5] Güera, D., & Delp, E. J. 2018. Deepfake Video Detection Using Recurrent Neural Networks IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) pp. 1–6

[6] Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020). Detecting Deep-Fake Videos from Appearance and Behavior.
IEEE International Workshop on Information Forensics and Security (WIFS) 1–6.

[7] Pallabi Saikia et al. 2022. "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features." *arXiv preprint arXiv:2208.00788*.

[8] Darius Afchar et al. 2018. "MesoNet: A Compact Facial Video Forgery Detection Network." *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7.

[9] Zhaohe Zhang & Qingzhong Liu (2020). "Detect Video Forgery by Performing Transfer Learning on Deep Neural Network." In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, Springer, pp. 387–396.

[10] W. Yassin et al. 2023. "An Integrated Deep Learning Deepfakes Detection Method (IDL-DDM)." In *High Performance Computing, Smart Devices and Networks*, Springer, pp. 81–91.

[11] Zhang et al. 2023. "An Enhanced Deep Learning-Based DeepFake Video Detection and Classification System." *Electronics*, 12(1), 87.

[12] Jaiswal et al. 2023. "An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection." *Applied Sciences*, 14(21), 9754.

[13] Jaiswal et al. 2023. "Convolutional Long Short-Term Memory-Based Approach for Deepfakes Detection from Videos." *Multimedia Tools and Applications*

[14] W. Yang, C. Hui, Z. Chen, J.-H. Xue, and Q. Liao .2019."Fvgan:finger vein representation using generative adversarial networks," IEEE Transactions on Information Forensics and Security, 14, 9, pp. 2512–2524.

[15] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, 2018 "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," IEEE transactions on pattern analysis and machine intelligence, 41, 8 1947–1962

[16] F. Liu, L. Jiao, and X. Tang,2019 "Task-oriented gan for polsar image classification and clustering," IEEE transactions on neural networks and learning systems, 30, 9, 2707–2719.

[17] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun,2019 "3d aided duet gans for multi-view face image synthesis," IEEE Transactions on Information Forensics and Security, 14, 8, 2028–2042.