Grenze International Journal of Engineering and Technology, June Issue

Enhancing Deepfake Detection: A Comparative Study of MesoNet, XceptionNet, EfficientNet, Optical flow and XceptionNet + Optical flow

Dr. Bhushan Yelure[1], Aditya Shingate[2], Altaf Mubarak[3], Aniket Bagal[4], Prathamesh Magare[5] and Saket Jadhav[6]
[1]Assistant Professor, Department of Information Technology, Government College of Engineering Karad, Satara, Maharashtra, India
Email: bhushan.yelure@gcekarad.ac.in
[2-6]Research Scholar, Department of Information Technology, Government College of Engineering Karad, Satara, Maharashtra, India
Email: {aditya.shingate, altaf.mubarak, aniket.bagal, prathamesh.magare, saket.jadhav}@gcekarad.ac.in

***Abstract*—Deepfake videos, which are manipulated to show false information, are becoming a major challenge in digital media. In this study, we explore four techniques—MesoNet, XceptionNet, Optical Flow, EfficientNet and an Ensemble Hybrid Model to detect deepfake videos. MesoNet is a lightweight neural network that focuses on detecting tampered images; XceptionNet is a more advanced model that identifies subtle fake features in videos, Optical Flow analyzes the movement between video frames to find inconsistencies that may indicate tampering, EfficientNet provides a balance between computational efficiency and accuracy by scaling neural networks effectively. We tested these models individually and in combination to determine the most effective approach for deepfake detection. By leveraging both spatial features (visual details) and temporal features (movement across frames), we achieved better identification of fake videos. This study highlights the importance of integrating spatial and temporal features along with different techniques to address growing threat of deepfakes and ensure the integrity of digital content in real-world applications.**

***Index Terms*— Deepfake, MesoNet, XceptionNet, EfficientNet and Optical flow.**

I. INTRODUCTION

Recent Recent breakthroughs and enhancements in Natural Language Processing (NLP), Computer Graphics, Optimization Techniques, and Machine Learning (ML) have enabled the creation of highly realistic audio, images, and videos where virtually anyone can be mocked to say or do anything. By collecting sufficient sample recordings, for instance, it's possible to replicate a person’s voice with remarkable accuracy. Similarly, with enough sample images, entirely fictional individuals can be created. Furthermore, realistic videos can be produced depicting anyone performing any action or speaking any dialogue the creator desires [1]. The use of sophisticated artificial intelligence models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), SimSwap and many more, the deepfakes enable users to alter text, audio, video, and image information in ways that are frequently indistinguishable to the human eye.

Grenze ID: 01.GIJET.11.2.585_1
© Grenze Scientific Society, 2025

This technology can produce hyper-realistic alterations, including face-swapping, voice synthesis, and image modification, which have been used in everything from entertainment to cybercrimes. While deepfakes open up new creative possibilities in media and advertising, they also pose serious ethical and security concerns. Notably, they are commonly used to generate fake pornographic content and political misinformation, which undermines public trust and can result in personal or societal harm. Document verification is an area of digital forensics that has long grappled with issues of authenticity, yet recent AI advancements have made it much more challenging to discern manipulated content. The United States defence agency initiated the Media Forensic (MediFor) project to support research on media integrity, emphasizing the need for digital, physical, and semantic confirmation of digital data. The ease of access to deepfake tools and web-based applications, such as FaceApp and DeepNude, has amplified the need for sophisticated detection methods to protect individuals and institutions from malicious use of these technologies. In response to these challenges, the study will provide a comprehensive overview on following: deepfake generation, detection techniques, highlighting the latest tools, datasets, and methodologies used to identify altered media. By offering an in-depth analysis of detection technologies and the datasets employed, we aim to support ongoing research and development in multimedia forensics. Key contributions of this study include a detailed examination of deepfake creation methods, taxonomy of detection techniques, and insights into current capabilities and limitations in combating this emerging threat. This work seeks to aid researchers, policymakers, and practitioners in better understanding and navigating the complex landscape of deepfakes, thereby helping to maintain trust and security in digital interactions [13].

II. Literature Survey

In response to these challenges, the study will provide a comprehensive overview on following: deepfake generation, detection. The fake videos created with deepfakes include Face2Face, Face-swapping, NeuralTextures [3], and Face-reenactment. Face-swapping is the replacement or swapping of one face in an image or video with another, so altering the identity of the person in the video. This method used two pairs of encoder-decoder, in which the both the encoders exchange parameters over training, while the decoders are trained separately. The Face-swapping method only function if there is a huge quantity of pictures and clips used as learning data, such methods where used to provide the deepfakes first. This initial face swapping approach can be traced back to 2017 on Reddit [1]. Deepfake technology utilizes advancements in ML, Deep Learning (DL), Artificial Intelligence (AI) and Computer Vision to make incredibly realistic fake content. Key enablers of this technology are GANs, Autoencoders, and Neural Networks. While these innovations present opportunities in entertainment and healthcare, they also raise ethical concerns, including privacy violations, misinformation, and national security threat [4] [5] [6].

There is wide range of Deepfake Creation Techniques such as Neural Networks, Autoencoders, GANs, Progressive GANs, etc. Variants like Deep Convolutional GANs (DCGANs) and Coupled GANs (Co-GANs) have been instrumental in improving output quality [4] [7]. The GANs is made up of two rival networks: the generator and the discriminator, which enhance their capabilities through an iterative process to create exceptionally lifelike results. The Neural Networks employ multi-layered neural networks for feature extraction, allowing precise replication of facial features and expressions [4]. The Autoencoders uses encoding-decoding structures for manipulating and synthesizing video or audio content [6] [8]. Progressive GANs was introduced by NVIDIA, these networks incrementally enhance resolution, enabling the generation of photorealistic images. The generation of deepfakes includes two contrasting viewpoints: the optimistic and the pessimistic. On the optimistic side, it can enhance movie effects, animation, and provide therapeutic solutions for speech or emotional recovery [4] [6].

On the pessimistic side, however, it can be used for the fabrication of misinformation, privacy intrusions, electoral manipulation, and pose threats to digital forensics and cybersecurity [12]. Detection methods for deepfakes have evolved to address the challenges posed by synthetic content across different media formats. Image-based detection algorithms focus on identifying Gaussian blur, noise artifacts, and pixel-level inconsistencies that are often present in deepfake images [9]. Video-based detection techniques analyze temporal artifacts across frames, using different variants of Neural Networks to improve accuracy in detecting manipulations [10] [11].

For speech deepfakes, new approaches have emerged that detect acoustic inconsistencies by analyzing spectral features using Deep Learning (DL) models. Additionally, hybrid approaches combine super-resolution algorithms and Residual Networks (ResNet) to expose subtle artifacts left by deepfake generation processes, further enhancing detection capabilities. These methods collectively aim to improve the identification of deepfake content and mitigate its potential misuse. Detecting deepfakes presents significant challenges due to the

increasing sophistication of GANs, which make it difficult to distinguish between real from synthetic contents [4]. Advanced techniques such as lip-syncing algorithms and high-resolution face-swapping further complicate the detection process, as they can create highly convincing and seamless manipulations [10]. Additionally, existing detection methods often struggle to generalize across different datasets, as the variability in training data and the diverse styles of deepfakes lead to inconsistencies in detection performance [6]. These challenges highlight the need for new innovative ideas in detection technologies to keep pace with advancements and to outmaneuver the deepfake generations. Preventive and legislative measures are being developed to combat the misuse of deepfake technology. Watermarking tools, which embed immutable metadata such as timestamps and location data into original media, can help track alterations and verify the authenticity of content [9]. Social media and streaming platforms are also implementing automated detection systems for instantly detecting and reducing deepfake content. Furthermore, legislative actions are being introduced to enforce laws that require the rapid removal of harmful deepfakes and impose penalties on creators of malicious content. These combined efforts aim to reduce the spread of deceptive media and protect individuals from the negative impacts of deepfakes. The dual-use nature of deepfake technology necessitates continuous innovation in both generative and detection domains. While advancements in GANs and neural networks fuel progress, robust detection and ethical frameworks are essential to mitigate societal risks. The proposed research addresses technical challenges while balancing innovation with responsible usage.

III. METHODOLOGY

A. Problem Definition

The motive of this research investigation is to improve and create broadly applicable model/framework for deepfake detection that can accurately classify images and video content as either authentic or manipulated.

- Challenges Addressed: Deepfake content often includes subtle yet convincing manipulations, such as altered facial expressions or texture inconsistencies that are difficult to detect with traditional methods. By addressing these nuances, the framework aims to mitigate the risks posed by synthetic media in areas such as misinformation and security.
- Multi-Domain Analysis: The system incorporates spatial, temporal, and frequency domain analyses to ensure comprehensive detection. This multi-modal approach provides a holistic view of the content, making the detection process more reliable.
- Input and Output Definition: The input comprises facial images or video sequences, while the output is a binary classification "real" or "deepfake". This classification forms the foundation for applications like media authentication and content moderation.

B. Data Collection and Pre-processing

Data for the study is sourced from publicly available datasets, including DeepFakeDetection (DFD), CelebDF, and a proprietary curated dataset, which collectively provide a diverse range of real and fake content. The DFD and CelebDF datasets are chosen due to their variety of manipulation techniques and high-quality labelled samples. Then in pre-processing we have series of following steps and they are Face Detection and Alignment to detect faces are aligned to ensure consistent orientation, Frame Sampling for Videos this step balances computational efficiency with the need to preserve temporal patterns for analysis, Normalization and Resizing of Image the reshaped image acts as a standardized the input for training a model and Data Augmentation is utilized to enhance the dataset by performing operations like flipping, rotating, contrast adjustment, and noise addition.

C. Feature Extraction

Feature extraction in this study is performed across three complementary domains: spatial, temporal, and frequency. Detection of Pixel-Level Artifacts is a Convolutional Neural Network (CNN) is used to identify inconsistencies in textures, lighting, and facial features that are indicative of deepfake manipulation. The spatial model focuses on capturing fine-grained details that differentiate authentic content from manipulated images. The Video Sequence Analysis with the help of Recurrent Neural Networks (RNNs) or Transformer architectures are employed to detect anomalies in video sequences.

The Movement and Transition Detection are Unnatural transitions, such as inconsistent blinking or abrupt changes in facial expressions, are identified to expose deepfake content are referred as Temporal Features. The Spectral Analysis can be done using techniques like Discrete Cosine Transform (DCT) and wavelet transforms are used to analyze image frequency components. Frequency domain analysis reveals hidden patterns, such as irregularities in spectral properties that are often introduced by deepfake generation algorithms.

D. Model Training and Ensemble Learning

The training process for deepfake detection involves developing individual models for spatial, temporal, and frequency features, which are then integrated into an ensemble framework. Spatial models, like CNNs, identify inconsistencies in image details, with transfer learning utilized to speed up training. Temporal models analyze video sequences to detect unnatural transitions, while frequency models focus on spectral inconsistencies. The predictions of the individual models are combined using a weighted voting strategy. The contribution of each model is assessed based on its effectiveness on the validation dataset. The ensemble approach ensures that spatial, temporal, and frequency features are utilized synergistically to improve detection accuracy. Five models are chosen based on their potential to captivate different aspects of the deepfake detection problem: XceptionNet, MesoNet, EfficientNet, Optical Flow, and a Hybrid Model combining XceptionNet + Optical Flow. MesoNet is a light-weight deep learning model, specifically developed for deepfake detection [14]. It also performs well even with smaller datasets, making it suitable for deepfake detection tasks where the available data may be limited or imbalanced. EfficientNet is a family of deep convolutional networks that optimize the balance between network depth, width, and resolution. It scales efficiently by using a compound scaling method that adjusts these factors proportionally, improving both performance and efficiency. XceptionNet is a deep CNN architecture based on depthwise separable convolutions, which divide ordinary convolutions into two pieces: depthwise convolutions and pointwise convolutions [15]. Due to its separable convolutions, XceptionNet can achieve high accuracy while being computationally efficient compared to traditional CNNs, which is critical when processing large video datasets. The depthwise separable convolutions allow the model to learn both low and high level features, making it capable of identifying intricate patterns that are typical in deepfake images. The motion pattern of objects between two successive video frames is commonly referred to as optical flow analysis, and it can be used to identify differences in object movement. Optical flow does not rely on pixel-wise labeling and instead focuses on frame-to-frame consistency, making it a powerful tool for detecting artifacts that are not genuine at the pixel level. The hybrid model combines the XceptionNet architecture with Optical Flow analysis, aiming to gather both spatial and temporal characteristics from videos. The hybrid model benefits from the complementary strengths of XceptionNet (spatial features) and optical flow (motion-based inconsistencies), making it more robust against various types of deepfake manipulation techniques. The fusion of spatial and temporal features often leads to superior performance compared to using either approaches alone, resulting in higher accuracy in detecting subtle artifacts that distinguish real from fake videos. Train, test, and validation sets make up the dataset. Training and testing sets are typically split 80-20. The training set consists of labelled real and fake video frames. Training setup also includes Loss Function in which Binary Cross-Entropy loss is used. Learning Rate prevents exceeding the minimum loss during training; a learning rate scheduler is used to modify the learning rate. The Batch Size is to strike a balance between model performance and computational economy, a batch size of 4 or 32 is selected. Depending on how well the training and validation losses converge, the model is trained for 20 epochs (refer Figure 6).

E. Performance Evaluation

The framework's performance is assessed using standard metrics along with their combinations and testing techniques. The evaluation metrics includes Accuracy denotes the proportion of accurately classified samples, Precision is the system’s ability to identify deepfake instances among predicted positives, Recall denotes the system's ability to identify all actual instances of deepfake within the dataset, F1-Score denotes the harmonic mean of precision and Recall, effectively balancing both to provide an evaluative metric (best between 0.8 to 1), providing a comprehensive metric for imbalanced datasets and ROC-AUC (Receiver Operating Characteristic Area Under the Curve) measures the trade-off between the true positive rate and false positive rate across various threshold values [2].

F. Testing and Validation

A separate test set is utilized to assess the model's capacity to generalize to novel and unidentified data. The performance analysis is done based on the examined results to identify areas of strength and potential adjustment. Insights from misclassified samples are used to refine the model and pre-processing steps.

IV. DEEP LEARNING MODEL

A. MesoNet

The deepfake videos exhibit unnatural transitions or inconsistencies. MesoNet, with its ability to learn from both spatial features (like faces, movements, etc.) and temporal dynamics (like smoothness of facial transitions and

continuity of movements), is capable of detecting these abnormalities. The input layer accepts a video sequence where each video is split into a series of frames. The shape is (10, 256, 256, 3) i.e. images (time_steps, height, width, channels) means 10 frames of 256 × 256 RGB. Since each frame in the video sequence is a separate image, we use TimeDistributed layers to apply the same CNN layers independently to each frame. This ensures that the CNN processes each frame of the video individually but shares the same weights. The Conv2D layer applies 16, 32 and 64 convolutional filters of size 3 × 3 over the input set, using the Rectified Linear Unit (ReLU) activation function and same padding, which retains the input dimensions while extracting feature maps. To reduce spatial dimensions, the max-pooling layers follow after each of the three convolutional layers [14]. Instead of being sent to the Dense layer directly, the data is flattened by being transformed into a one-dimensional vector. The dense layer captures the temporal dependencies between the frames. Also the Fully Connected layers process the aggregated temporal features and produce the final classification decision. The Figure 1 depicts the flow of MesoNet architecture followed in this process. A Dropout layer can be added for regularization to avoid overfitting. The sigmoid activation function, the last layer, determines whether the video is authentic (0) or fraudulent (1) by returning a value between 0 and 1.

B. XceptionNet

XceptionNet also known as Extreme Inception, a deep learning architecture primarily used for classification and other vision tasks. The shape of image is (299, 299, 3) which represents the image's width and height are represented by 299 × 299, and its three colour channels Red, Green, and Blue. The number of outcomes is two as we are performing binary classification (real vs fake), so this is set to 1, and the output will be a single probability (between 0 and 1) [15]. XceptionNet is a CNN-based deep learning model pre-trained on the ImageNet dataset, utilized for feature extraction in tasks like deepfake detection. The Figure 2 depicts the flow of XceptionNet architecture followed in this process. The model's convolutions are frozen to prevent retraining, leveraging transfer learning. Custom layers, including Global Average Pooling (GAP), Flatten, and Dense layers (128 and 64 neurons), help reduce overfitting and learn complex patterns. ReLU is used for non-linearity, and the final output layer uses a sigmoid activation for binary classification. The model uses a learning rate of 1e-4 and binary cross-entropy loss. EfficientNet optimizes CNN performance with compound scaling and MBConv blocks, improving accuracy and efficiency.

C. EfficientNet

It represents a significant advancement in CNN [16] use compound scaling to optimize the ratio of depth, width, and resolution. This allows it to achieve better accuracy and efficiency than traditional models, making it one of the most powerful and efficient architectures for a variety of deep learning applications. By employing techniques like MBConv blocks, Swish activations, and compound scaling, EfficientNet delivers an optimal solution for both high performance and low computational cost. The images (or frames from a video) are passed into the model The resolution of the input images is critical, EfficientNet can scale well to different image resolutions (224 × 224), and can work efficiently even on high-resolution inputs. EfficientNet begins with an initial stem layer, which typically consists of a convolution operation that processes the input image; this is where input image is transformed in a set of feature maps. The core of EfficientNet consists of multiple stacked MBConv blocks. Each MBConv block uses depthwise separable convolutions that separate the filtering of each input channel (depthwise convolution) from the linear combination of those filtered channels (pointwise convolution). This reduces the computational cost and allowing model to explore and learn more complex representations. The MBConv blocks also use swish activations, which are particularly beneficial for training deep networks and helps achieve better performance. It includes skip connections (residual connections) in the architecture, which help with the flow of gradients during training and make the network more efficient in learning complex representations. These connections improve model convergence and help prevent overfitting. This all is included in EfficientNet pertained and then passed to GAP to reduce the activation maps into a constant size vector (the resultant of this layer is a single value per feature map). GAP reduces the number of parameters and focuses on global features rather than local ones, which is particularly helpful for detecting inconsistencies across the whole image. The outcome of the GAP layer is sent via one or more utterly linked layers to map the high level features into class predictions. Then it is passed to dense layer having 128 and 64 neurons and then sends to final layer accommodating a sigmoid activation function for the final prediction.

D. Optical flow

It describes the apparent motion pattern of objects in a visual image that results from the relative motion of the scene and the observer (camera) [17]. In simple terms, optical flow refers to the variation in the speeds of

brightness patterns within an image, illustrating how objects seem to shift over time as the observer moves or the scene evolves. It's a key concept in computer vision and motion analysis. Optical flow detects pixel motion between consecutive frames using spatial gradients (Ix, Iy) and temporal derivatives. The video is split into frames of shape (64, 64, 3), with each frame representing x and y flow components and the temporal change.

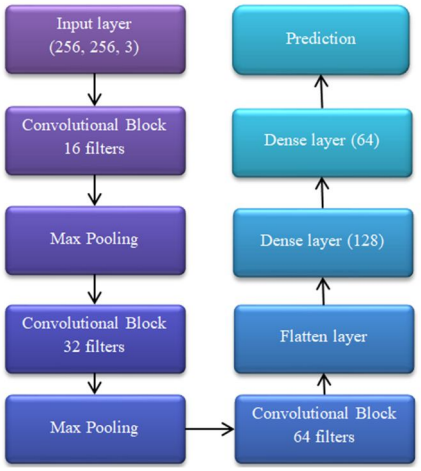

Figure 1. MesoNet Model

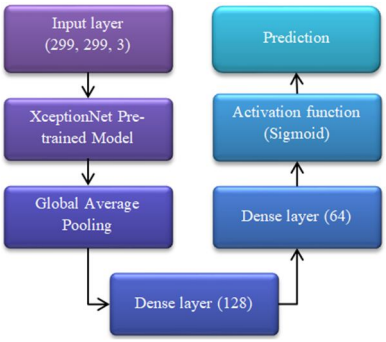

Figure 2. XceptionNet Model

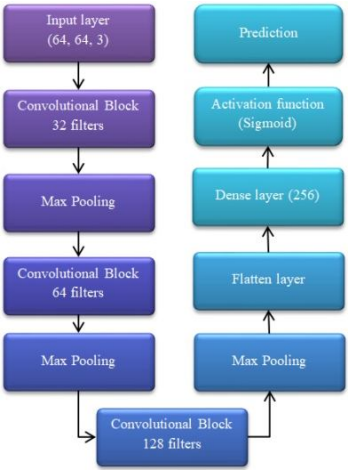

Figure 3. Optical flow Model

The frames are resized, normalized, and augmented for robustness. Optical flow algorithms like Lucas-Kanade or Horn-Schunck compute the motion vector field. Dense optical flow captures motion for all pixels, while sparse flow focuses on specific points (e.g., facial landmarks). The model uses three convolutional layers with max-pooling, followed by a dense layer with 256 neurons, and a sigmoid activation for binary classification. Dropout reduces overfitting, and Adam optimizer with binary cross-entropy loss ensures efficient training. The Figure 3 depicts the flow of Optical flow architecture followed in this process.

$$I_x u + I_y v + I_t = 0 \qquad (1)$$

E. XceptionNet + Optical flow

The proposed hybrid approach for identifying deepfakes combines the extraction of temporal and spatial features. The optical flow calculation captures motion information by computing the flow between consecutive video frames using OpenCV's Farneback method, converting it to magnitude and angle to represent motion intensity and direction.

The data class loads labelled video frames from "real" and "fake" folders, compute optical flow for each frame, and apply transformations like resizing and augmentation. The model architecture uses Xception to extract spatial features from video frames, while a custom CNN processes optical flow images to capture temporal features. These characteristic features are combined and handover to connected layers for classification. During training of model we employed cross-entropy loss alongside Adam optimizer to optimize the model, with accuracy tracked over epochs. The parameters of the model are stored, and its performance is assessed utilizing accuracy and a confusion matrix on a test dataset. This hybrid approach leverages both appearance and motion cues for more robust deepfake detection. The Figure 4 depicts the flow of hybrid architecture followed in this process.

V. DATASET

The dataset for this deepfake detection contains a total of 2,200 videos, with an equal distribution of 1,100 fake videos and 1,100 real videos. The fake videos in the dataset represent a diverse range of deepfake generation techniques, including Deepfakes, Face2Face, FaceSwap, and NeuralTextures. These methods vary in their approach to face manipulation, making the dataset particularly challenging for deepfake detection algorithms. Deepfakes utilize deep learning models to generate highly realistic but synthetic faces, while Face2Face and FaceSwap involve swapping faces in real-time from one video to another. NeuralTextures, another advanced technique, generates realistic textures and facial movements, further increasing the difficulty of detection. The mix of these techniques ensures that the dataset is comprehensive and represents the landscape of evolving deepfake technology, providing a robust foundation for detection models.

TABLE I. TRAINING ACCURACY

	MesoNet	XceptionNet	EfficientNet	Optical flow	XceptionNet + Optical flow
5 epochs	87.70	88.97	85.02	87.23	80.46
10 epochs	88.31	89.05	85.46	89.91	83.24
15 epochs	86.56	89.71	85.56	90.63	89.45
20 epochs	89.80	89.97	85.58	93.45	89.37

TABLE II. TESTING ACCURACY

	MesoNet	XceptionNet	EfficientNet	Optical flow	XceptionNet + Optical flow
Accuracy	88.19	88.44	82.85	85.75	85.87

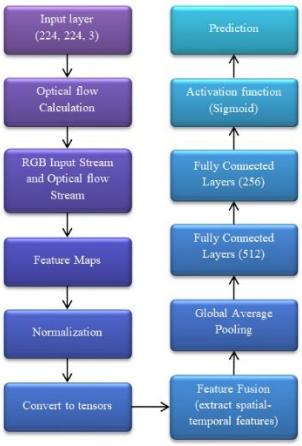

Figure 4. Optical flow + XceptionNet (Hybrid) Model

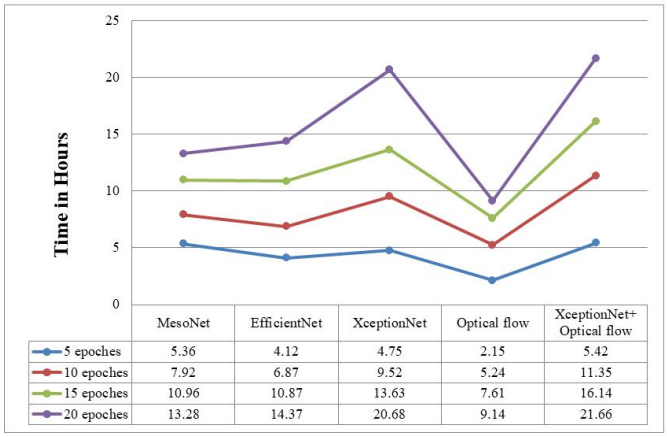

Figure 5. Loss Function

VI. DISCUSSION AND EXPERIMENTAL RESULT

In this study, several deepfake detection models were evaluated for their effectiveness in identifying synthetic media. The models tested include MesoNet, XceptionNet, Optical Flow, and EfficientNet, each contributing differently to the detection task. MesoNet achieved an accuracy of 88.09%, demonstrating solid performance in detecting deepfakes. This model, known for its efficiency in processing facial features, is relatively lightweight compared to more complex networks, yet still offers good detection results. XceptionNet, with an accuracy of 89.42%, outperformed MesoNet, benefiting from its deeper architecture and use of depthwise separable convolutions. Its higher performance suggests it can capture more nuanced patterns in the data, making it more effective for deepfake detection. Optical Flow, with an accuracy of 90.31%, performed the best among the tested models which can be seen in Table I. By analyzing motion patterns between frames, Optical Flow was able to detect inconsistencies in the dynamic aspects of deepfakes, resulting in superior performance among all the other models. EfficientNet, achieving 85.40% accuracy, showed a low performance relative to the other models. While EfficientNet is known for its efficiency in resource usage, it struggled with deepfake detection, likely due to the challenges of capturing fine-grained facial or motion anomalies inherent in synthetic media. The Xception+Optical Flow hybrid model, with an accuracy of 86.73%, demonstrated a solid performance in deepfake detection. By combining Xception's powerful feature extraction capabilities with Optical Flow's ability to track motion patterns across frames, the model improved its ability to capture subtle facial and motion discrepancies that are characteristic of deepfakes. However, while the model outperformed EfficientNet, its performance still indicates room for further optimization, particularly in handling extreme variations in synthetic media and complex facial distortions. All the models have high initial building time as epoch’s progress the time reduces to a certain interval. The hybrid model takes 21.66 hours to train, while optical flow takes 9.14 hours to train and the average overall training time is 15.83 hours can we seen in Figure 5. We can use high computational resources to shorten the training duration.

VII. CONCLUSION

In conclusion, the evaluation of several deepfake detection models revealed varying levels of performance, with Optical Flow emerging as the top performer at 90.31% accuracy (refer Figure 7). This model's ability to analyze

motion discrepancies between frames proved highly effective in detecting the dynamic inconsistencies typical of deepfakes. Due to the deeper design and capacity to identify more complex patterns in the data, XceptionNet outperformed MesoNet in terms of performance as well. MesoNet, while lightweight, delivered solid results with an accuracy of 88.09%, showcasing its efficiency in detecting facial features. EfficientNet, though known for its resource efficiency, struggled with deepfake detection, achieving the lowest accuracy at 85.40%. The Xception+Optical Flow hybrid model, with an accuracy of 86.73% while training and 85.83 on testing (refer Table II), offered a promising combination of feature extraction and motion tracking, though it still leaves room for improvement, especially in dealing with complex facial anomalies and more extreme variations of synthetic media. Overall, the study emphasizes how crucial it is to include both spatial and temporal information for reliable deepfake detection. It also emphasizes the necessity of additional optimization to meet the problems presented by increasingly complex deepfake algorithms.

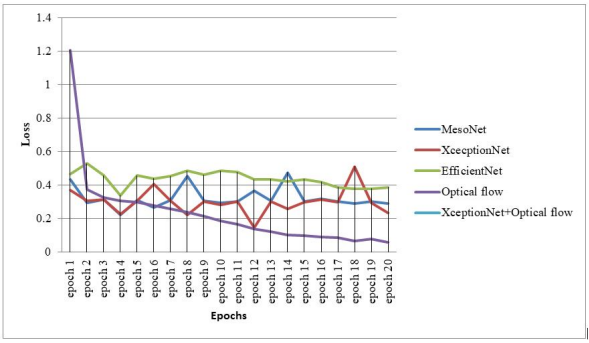

Figure 6. Loss function

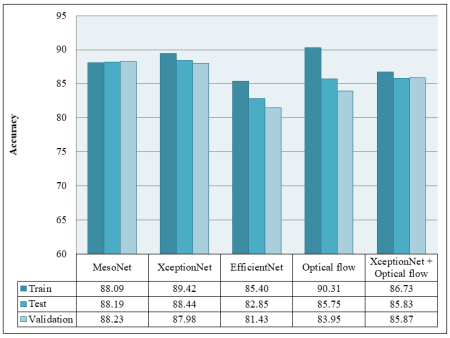

Figure 7. Train, Test and Validation Accuracy

REFERENCES

[1] Shruti Agarwal, Hany Farid, Ohad Fried, Maneesh Agrawala; Deep-Fake Videos from Phoneme-Viseme Mismatches, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 660-661

[2] Deepfake Detection through Deep Learning Chawla, R. (2019). Deepfakes: How a pervert shook the world. International Journal of Advance Research and Development, 4(6), 4-8.

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Niessner, FaceForensics++: Learning to Detect Manipulated Facial Image, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11

[4] Negi, S., Jayachandran, M., & Upadhyay, S. (2021). Deepfake: An Understanding of Fake Images and Videos. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology.

[5] Nguyen, T. T., et al. (2019). Deep Learning for Deepfakes Creation and Detection. arXiv preprint.

[6] Tolosana, R., et al. (2020). Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion.

[7] Jonathan Hui (2018). GAN - DCGAN (Deep Convolutional Generative Adversarial Networks). Blog.

[8] Felix Mohar (2017). Implementing a Generative Adversarial Network to Draw Human Faces. Towards Data Science.

[9] Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake Detection by Analyzing Convolutional Traces. CVPR Workshops.

[10] Li, Y., et al. (2020). Celeb-DF: A Large-Scale Dataset for Deepfake Forensics. CVPR Proceedings.

[11] Koopman, M., et al. (2018). Detection of Deepfake Video Manipulation. Irish Machine Vision and Image Processing Conference.

[12] Korshunov, P., & Marcel, S. (2018). Deepfakes: A New Threat to Face Recognition? Assessment and Detection. arXiv preprint.

[13] Ivanov, N. S., et al. (2020). Combining Deep Learning and Super-Resolution Algorithms for Deepfake Detection. IEEE.

[14] Darius Afchar, Vincent Nozick, Junichi Yamagishi, Isao Echizen; MesoNet: a Compact Facial Video Forgery Detection Network, 2018 IEEE International Workshop on Information Forensics and Security (WIFS).

[15] Francois Chollet; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251-1258, Xception: Deep Learning With Depthwise Separable Convolutions

[16] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML 2019). https://arxiv.org/abs/1905.11946

[17] Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo, Deepfake Video Detection through Optical Flow based CNN