

AI BASED DEEPPAKE DETECTION TECHNIQUES

Mohammed Hamza, Tameem Mansoor

Jain Deemed to be University, Bangalore, India

Abstract—Recent progress in deepfake algorithms has allowed visually compelling deepfakes to be generated from raw pixels with limited visual artifacts. Their applications range from positive use cases such as media creation to nefarious uses such as misinformation and media-based frauds. With state-of-the-art image generation models becoming photorealistic, identifying synthetic media is also becoming more difficult [14]. We introduce a multi-stage deepfake detection framework that leverages deepfake detection using spatial domain representations, frequency-domain representations, and temporal modeling of frames to produce a more accurate and robust framework. We incorporate a multi-scale EfficientNet [1] backbone for spatial artifact detection, a frequency-domain branch that learns inconsistencies introduced during the synthesis process, and a temporal encoder that models relationships at the frame-level with a Transformer [2] backbone. Models were trained on FaceForensics++ [5] and tested on Celeb-DF v2 [3], DeepFake Detection Challenge (DFDC) [4] datasets to analyze cross dataset generalization. The proposed multi-branch framework reached 99.8% accuracy on Celeb-DF dataset and 97.6% on DFDC, outperforming CNN baseline and CNN-Transformer hybrid. Generalization was also measured with adversarial attacks created using Fast Gradient Sign Method (FGSM). Proposed architecture exhibited significantly higher accuracy when subjected to adversarial attacks. An additional user awareness survey of 70 participants was performed to gauge public opinion on deepfakes and their trust in automated detection. Results show improvement in accuracy and robustness when combining spatial, frequency-domain and temporal representations of data for deepfake detection across datasets

Keywords—Deepfake detection, Deepfake algorithms, Synthetic media, Photorealistic image generation, Misinformation, Media-based fraud

I. INTRODUCTION

Deep learning has made it much easier to generate convincing fake media. New kinds of neural rendering, like Generative Adversarial Networks (GANs), allow us to generate videos and images that look very realistic. These advancements if used for good can provide substantial value to industries such as film production, virtual reality and digital content creation; however, they also provide an increased ability to generate maliciously manipulated content (known as Deepfakes) and propagate this content, with the intention of committing identity theft, financial fraud and/or political manipulation. This creates real risks to the integrity of digital security and to the trust people have in the accuracy of content found online.

Social media has been used to disseminate manipulated media, resulting in ongoing concerns surrounding misinformation, identity theft and political manipulation; thus, emphasizing the need to develop reliable automated deepfake detection systems that can be used to verify digital media.

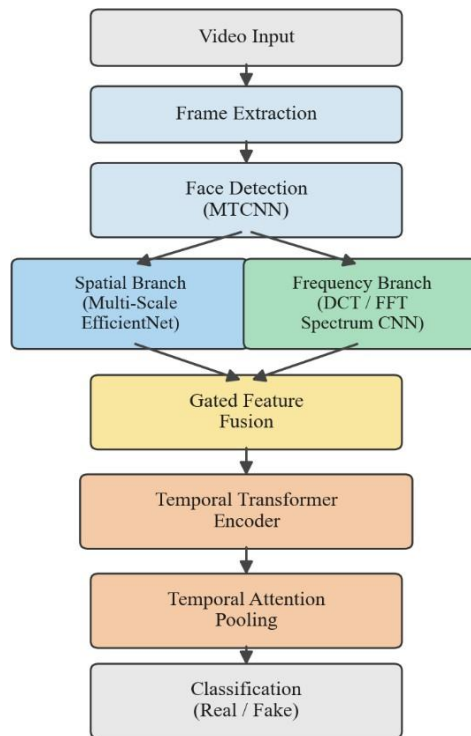
Recent studies show that deepfake generation methods are changing really fast, making it harder to find deepfakes. Most detection techniques in the beginning focused on finding visual problems or other inconsistent components in the modified images. However, the improved quality of generative models means that those visual problems are becoming less obvious to see, and can't be spotted as easily with standard computer vision detection systems. Additionally, public datasets like FaceForensics++ and Celeb-DF have been created to help with research and provide large, pre-recorded video samples of real and deepfake videos to allow for training and testing purposes [3], [4]. Overall, despite advancements made in the detection of deepfakes, many of the detection models trained on one dataset still struggle when tested on any other datasets or are still susceptible to adversarial manipulations of those models after they have been built.

The three features for detecting deepfakes are spatial artifacts in isolated frames, temporal differences in the sequences, and frequency domain characteristics from image synthesis. CNNs are used to detect spatial artifacts in manipulated frames, while RNNs and other sequence learning algorithms extract temporal differences in facial motion and blinking patterns [6]. Recently, Transformer-based architectures have been shown to be powerful for accurately modeling long-range dependencies within sequential information [8]. Prior research indicates that most deepfake image synthesizers produce fingerprints in the frequency domain that allow for detection because of the inconsistency created during the image creation process [11]. By combining and using complementary feature representations, there exists an opportunity for increasing the dependability of deepfake detection systems.

This paper's proposed multi-stage deepfake detection framework will improve the accuracy and robustness of deepfake detection by using spatial, frequency-domain and temporal representations. It uses a multi-scale EfficientNet backbone for

a spatial feature extraction branch, frequency domain analysis branch for detecting spectral artifacts and Transformer-based temporal encoder for modeling temporal (frame-level) dependencies in videos. In addition to testing how well this framework performs on benchmark datasets, the robustness of the proposed model is assessed utilizing adversarial perturbations created with Fast Gradient Sign Method (FGSM) [1]. In addition, an online survey will be conducted to gather information from the general public about their perceptions of deepfakes and their need for dependable automated detection tools.

The general structure of the overall proposed deepfake detection framework is shown in Figure 1, including preprocessing, both branches of feature extraction — spatial and frequency — and using a Transformer encoder for temporal modeling and final output classification.



The important things about this study are as follows:

1. A deepfake detection framework that is made of many things needed to find deepfakes by using different methods (spatial, frequency domain, and temporal feature representation).
2. A study of other methods or hints to help find deepfakes (artifacts that can help tell whether the picture is real; predicting disagreements across datasets; possibly amplifying or reducing the artifacts).
3. A way to evaluate deepfake performance by putting together datasets that can assess or generalize how well a new dataset will perform when tested with previous datasets (using three different data sources to evaluate the frameworks outlined in 1 and 2; how well does each of the three different proposed techniques work when a new dataset is introduced).
4. An evaluation of the attacks against these frameworks by conducting FGSM experiments to see how well they can resist sophisticated attacks against them (to determine what improvements can be employed in developing new frameworks).
5. A study involving individuals to understand their current mindset towards deepfake technologies and liability to use automated detection techniques (i.e., how do they view deepfake technology and is it reliable enough to make the correct decision?).

This chapter is organised as follows: Chapter 2 reviews the literature on generating and detecting deepfakes. Chapter 3 presents the proposed framework for detecting deepfakes. Chapter 4 describes the experimental data used for this study and how the experiments will be completed. Chapter 5 contains a report of the experimental results and analysis. Chapter 6 discusses the subject of the user awareness study (the report on user perception of deepfake technology). Finally, Chapter 7 contains the conclusions from the study and offers future recommendations.

II. RELATED WORKS

2.1 Deepfake Generation Techniques

Deep learning continues to change at an exponential rate and developing realistic synthetic media using generative models is one of the most rapidly developing technologies. One of the most important of these technologies would be the use of Generative Adversarial Networks (GANs) to create and/or identify visually real images and video. GANs are essentially made up of two Neural Networks known as a generator and a discriminator that compete against each other in order to produce visually realistic content that can pass as real. Due to the rapid advancements of the technology behind generative modelling we are now increasingly finding it difficult to separate deepfake content from legitimate media. Legitimate uses of generative modelling technology include face swapping, facial re-enactment, and neural rendering; however, the same technology is also being used in the commission of malicious acts such as creating false news, impersonating someone's identity, and digital financial fraud. Therefore, the need for accurate methods to detect deepfake content represents an important new area of study in both the fields of computer vision and digital forensics.

In this domain there are numerous benchmark datasets available to assist in conducting research. The FaceForensics++ dataset contains many manipulated videos created with various deepfake methods which allow researchers to evaluate different detection models in a controlled environment [3]. The Celeb-DF dataset was created to provide detection tasks that are quite difficult due to more realistic deepfakes being produced since they contain fewer visual artifacts compared to their counterparts [4]. In addition, the DeepFake Detection Challenge (DFDC) dataset was established in order to create a large-scale benchmark for evaluating the performance of different detection approaches against a wide range of video sources.

2.2 Deepfake Detection Approaches

Over the years, various methods for detecting deepfake videos have been developed, with developments evolving from purely identifying spatial anomalies in single frames to using temporal, or time-based, approaches for identifying anomalies throughout a complete video.

Historically, methods to detect deepfakes used CNNs, which are neural networks that can learn to identify patterns in images, as the baseline for identifying spatial anomalies in a video. CNN models analyze all frames of a video for inconsistencies in facial textures, lighting, and edges, especially where there may be subtle indicators of tampering in the manipulated data.

CNN models that have been designed for these tasks, such as EfficientNet, have produced high detection rates while also being computationally efficient enough to be effectively used in deepfake detection systems.

Since spatial approaches alone are not very accurate at identifying deepfakes, many researchers have also used temporal approaches, such as using recurrent neural networks and sequential models – which are neural network designs that can identify temporal anomalies, such as irregular motion patterns, unnatural blinking patterns, and imperfect lip-syncing in manipulated videos.

Recently, researchers have also been exploring attention-based models to improve deepfake detection rates using temporal relationships in video sequences. Because the attention mechanism built into those models allows them to identify temporal inconsistencies between disassociated frames in a video, using them greatly enhances the ability to identify the presence of temporal inconsistency in a deepfake video.

2.3 Frequency-Domain and Artifact-Based Detection

Further exploration into Deepfake artifact detection, researchers have investigated frequency-domain representations, in addition to temporal and spatial attributes. Artifact creation via Generative Models frequently generates spectral inconsistencies through their generative synthesis process. These inconsistencies can be detected by analysing image data in the frequency-domain. Research data have established that distinctive frequency-based signatures or fingerprints exist for images and videos created with deepfake frauds [11]. These signatures are attributed to the creation process and can only be identified through examination of frequency-based representation of images using either the Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT).

The use of spatial and frequency-based features together has been demonstrated to enhance the robustness of deepfake detection models. This is particularly evident when visual clues have diminished as a by-product of better quality being generated from the generative model. Therefore, combining both spatial and frequency-based features provides the most extensive and comprehensive technique for detecting false media.

2.4 Adversarial Vulnerabilities in Detection Systems

Deepfake detection algorithms have made considerable advances over the last few years; however, they are still susceptible to attacks made using adversarial examples. Attackers manipulate an input sample by adding slight modification to the data and produce a false prediction by passing it through a machine learning model. One of the most common and earliest techniques used in generating adversarial examples is the Fast Gradient Sign Method (FGSM) [1]. The influence of adversarial manipulation on deep learning models for image classification and detection demonstrate how critical it is to evaluate model robustness to different types of adversarial attacks [12].

These vulnerabilities cause challenges when deploying deepfake detection systems in real-world environments where adversaries will try to avoid detection using automated detection systems. As such, assessing detection models' performance against adversarially generated data has become a fundamental part of current deepfake detection research.

Even though there are countless detection techniques based on spatial, temporal, and frequency features, many of them fail to achieve optimal performance when evaluated on various datasets and against various types of adversarial attacks. Therefore, there exists a need for improved detection standards that implement multiple types of features into detection frameworks and evaluate their performance across all aspects of experimentations.

III. PROPOSED DEEFAKE DETECTION FRAMEWORK

We will present an outline of the deepfake detection system, which includes video analysis of manipulated images through different forms of representation in the spatial frequency domain and temporal features. The overall framework is made up of multiple detection components to increase accuracy and improve the robustness of detection across many datasets. The framework consists of a preprocessing step for videos and face extraction; a multi-scale representation of spatial features through feature learning; a detection from frequency-domain artefacts; and a temporal model for analysing a sequence of images through a transformer-based encoder.

3.1 Data Preprocessing

Prior to being processed by the detection framework, video data must go through a number of preprocessing steps. Each video sequence is first separated into individual frames. Each frame is then put through a face detection process with the Multi-task Cascaded Convolutional Network (MTCNN) for face detection followed by the cropping of the detected faces and resizing them to a constant size of 224×224 pixels (this is done to have uniformity across all of the training samples).

To help improve the model's ability to generalize, a number of augmentation techniques will also be used during preprocessing. Augmentation techniques can include horizontal flips, Gaussian noise, and JPEG compression artifacts which all represent different types of distortion that can occur in real-world videos. The use of augmentations will provide an opportunity for the model to learn robust feature representations that will be applicable across a wide variety of imaging conditions.

3.2 Multi-Scale Spatial Feature Extraction

An EfficientNet-based convolutional backbone is used as the basis for spatial artifact detection. EfficientNet has been shown to be an effective method for image classification due to its ability to balance the scaling of depth, width and resolution of the network, while being computationally efficient [9]. As part of this framework, features will be extracted from multiple intermediate layers of the EfficientNet architecture in order to capture artifacts at multiple spatial scales.

Deepfake generation methods create subtle variations in facial texture, region of boundary, and illumination pattern. By aggregating feature maps from different levels of the EfficientNet architecture, fine-grained pixel-level artifacts and larger structure inconsistencies in the manipulated face region can both be identified. The multi-scale representation will be projected to a common embedding space and fused together by attention.

3.3 Frequency-Domain Feature Analysis

A growing body of research suggests that generative models typically exhibit recognizable characteristics in the frequency domain by producing images that contain artifacts that are not visible in the spatial domain but can be detected using spectral analysis methods [11].

The frequency analysis branch in this study involves the use of transformations such as the Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT), which provide a spectral representation of frame representations by converting spatial image representations into spectral image representations.

Spectral representations highlight periodic patterns and frequency irregularities that can be seen in images generated by digital manipulation. The extracted frequency features from the convolutional features will then be combined with the spatial features by utilizing a feature fusion layer.

3.4 Temporal Transformer Modeling

Detecting deepfake videos requires analyzing relationships across frames rather than just the features in single frames. When analyzing video content, the most significant form of evidence used to detect such manipulation is by determining the temporal inconsistencies found throughout the video, such as facial motion not matching the video content or unnatural blinking patterns or not achieving lip synchronization. Previous work in detecting deepfake videos used recurrent neural networks (RNNs) which attempted to understand how temporal dependencies were created [6]. Due to their limitations when attempting to model long-range dependencies in long video sequences using RNNs, the suggested framework will include a Transformer-based temporal encoder, being inspired by the self-attention mechanism introduced by Vaswani et al. in 2017 [8]. The Transformer processes multiple frame-level features embedding sequences and adds a positional encoding to maintain the temporal order. The self-attention mechanism allows the model to determine what relationship exists between frames that are further apart in the sequence, assisting in improving detection capability for a wide variety of subtle temporal discrepancies within the video sequence.

3.5 Hybrid Feature Fusion and Classification

The framework combines spatial, frequency, and temporal features using a hybrid fusion method after they are extracted. A gated fusion module creates a weight for how important the spatial representation provides to the frequency representation within each dimension of feature space. The resultant feature sequence from the fusion module is passed as input to the temporal Transformer encoder which produces a contextualized feature embedding for the entire video.

Finally, a temporal attention pooling layer aggregates the resulting sequence representation into one final feature values that will be input to the binary classification layer. The classification layer predicts whether the input (source) video is real or falsified (streamed). The proposed framework's combination of spatial artifact detection with frequency-domain analysis and temporal modeling creates a complete method for detecting manipulated media across various datasets and adverse environmental conditions.

3.6 Exploratory Detection Signals

In addition to the main detection pipeline, the developer of this framework conducted some brief explorations.

Investigations of the exploratory detection signals. The purpose of these signals was to determine if non-standard spatial and temporal representations of complementary signals could help to identify manipulated media. Signal 1, Artifact Consistency Analysis (ACA), looks at the "stability" across consecutive frames of the spectral artifact patterns. Real videos show a high degree of stability in their spectral characteristics, whereas the deepfake generation process might introduce subtle temporal instability into the distribution of those artifacts. Signal 2, Cross-Domain Prediction Disagreement, looks for variance among predictions from the spatial detection branch, frequency detection branch, and temporal detection branch. Large amounts of disagreement among the three independent domains may indicate a change in the characteristics of the manipulated media. Signal 3, Frequency Artifact Amplification (FAA), also uses high-frequency emphasis in the spectral domain to make subtle generative artifacts more prominent and detectable, even when they are visually present but hard to detect in the spatial domain. These three exploratory signals were implemented as lightweight auxiliary analyses during the developer's initial experimentation. Preliminary findings from these investigations suggest that they may provide useful complementary information for deepfake detection but a full evaluation of any of these three techniques is reserved for a later study.

IV. EXPERIMENTAL SETUP

This section describes the datasets, training configuration, evaluation metrics, and adversarial testing procedures used to evaluate the proposed deepfake detection framework. The experiments were designed to assess both detection accuracy and model robustness across multiple benchmark datasets.

4.1 Datasets

The FaceForensics++ data set is a large scale, manipulated facial video data set that has been generated using a few different deepfake generation techniques

such as DeepFakes, FaceSwap, and NeuralTextures. The dataset contains manipulated and original videos, and is often used as a benchmark for both training and evaluating deepfake detection systems [3]. FaceForensics++ was selected as the primary training dataset due to its multiple different manipulation techniques.

Celeb-DF v2 is a more difficult dataset which has been created specifically to improve upon the previous datasets, and includes high quality deepfake videos with significantly reduced visible artefacts [4]. The high quality of the videos in this dataset creates a much more challenging task for deepfake detection, and therefore a much higher quality benchmark for assessing model performance.

The DFDC dataset is one of the largest publicly available deepfake datasets and contains thousands of manipulated videos that have been created using a wide range of different synthesis techniques. The DFDC dataset was created as part of the DeepFake Detection Challenge and is widely used to assess the generalization capabilities of deepfake detection models across multiple video conditions.

In this experimental setup, the models were trained on the FaceForensics++ dataset and evaluated on the Celeb-DF v2 and DFDC datasets to determine cross-dataset generalization performance.

4.2 Training Configuration

Each model was built using the Pytorch deep learning framework. I utilized the Adam optimizer to train the models. The starting learning rate was 1×10^{-4} with a batch size of 32. The models were trained over 30 epochs employing binary cross entropy (BCEwithLogitsLoss), a loss function most often used for binary classification problems.

Using a learning rate scheduler and cosine annealing (a gradual decrease in the learning rate to provide more stable convergence), I was able to reduce the amount of time spent on training while increasing the stability of the models.

Using a mixed precision approach to training and using gradient clipping, I improved on the efficiency of my computation and the amount of memory used by my GPU through efficient model architecture design. While I reviewed the results of the object detection models that I created, I also considered the computational efficiency of each architecture as well. By combining a convolution-based EfficientNet network backbone with a Transformer network for temporal representation, I achieved a balance between the representational capacity of an object detector and computational efficiency.

Mixed-precision training and optimized data preprocessing have been employed to enhance computational speed of training stability and reduce memory footprint. Therefore, other factors impacting training can provide the necessary computational efficiency while producing appropriate results with high detection performance over a variety of datasets.

Besides the established main training path, several explore signal extraction modules were developed to explore various alternative techniques for capturing additional cues that could assist in detecting deepfakes. However, these will not be included in the final evaluation path.

4.3 Evaluation Metrics

To fully examine how well the proposed deepfake detection framework would work, standard classification performance metrics are employed: accuracy, precision, recall, F1-score and area under the receiver operating characteristic curve (ROC-AUC).

Accuracy is the total number of correctly classified samples. Precision is the total number of predicted manipulated videos that were recorded as manipulated; recall is the total number of actual manipulated videos that were identified correctly. The F1-score is a harmonic mean of precision and recall giving an overall balanced assessment of classification accuracy.

The ROC curve and corresponding area under the curve (AUC) were also used to assess the models' ability to differentiate between real and fake videos across different decision thresholds.

4.4 Cross-Dataset Evaluation

To assess the ability of the suggested detection framework to generalize, we ran experiments to train models using the FaceForensics++ dataset that could then be tested using the Celeb-DF v2 and DFDC datasets.

Cross-dataset evaluation is critical for determining how easily deepfake detection models created using one dataset will work on other datasets where the original dataset had never been trained with any of the manipulated images. Further, by performing cross-dataset evaluation across multiple datasets with varying manipulation characteristics, we can evaluate how robustly and appropriately the suggested solution can be used in the real-world.

4.5 Adversarial Robustness Testing

Alongside evaluating standard performance, we also tested the robustness of our proposed detection framework to adversarial perturbations. Adversarial attacks are used to create small perturbations that are carefully designed to mislead models into making wrong predictions.

We took multiple perturbation strengths and evaluated them to see how detection performance performs as adversarial noise increases. This evaluation will demonstrate if our proposed detection framework is resilient enough not only in an adversarial environment but also when malicious actors may have intentionally attempted to avoid being detected by our automated detection systems.

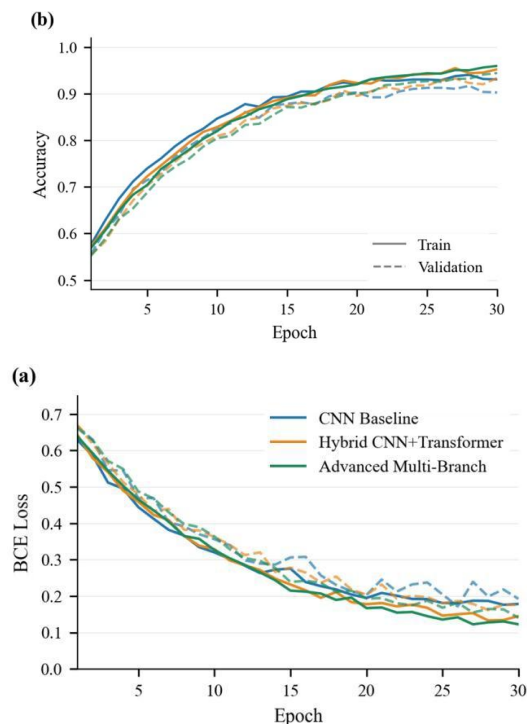
V. EXPERIMENTAL RESULT AND ANALYSIS

In this section, we present the experimental results produced by the proposed deepfake detection framework. The proposed models were evaluated using multiple datasets to determine their accuracy in detecting deepfakes, generalizing across multiple datasets and how well they resist adversarial perturbations. Three different configurations of the models were evaluated: an EfficientNet-based Convolutional Neural Network (CNN) baseline model; a hybrid CNN-Transformer temporal model; and an advanced multi-branch architecture that integrates spatial, frequency, and temporal features using multiple deep learning techniques.

5.1 Performance During Training

During training, we monitored each of the model's using loss and accuracy curves to check for signs of convergence and learning stability throughout the training process. All of the models demonstrated consistent signs of convergence throughout their respective training runs, with the hybrid and advanced model architectures demonstrating faster convergence times and lower final training losses compared to the baseline CNN model.

Figure 2 shows the training loss and accuracy curves for the various architected models evaluated during their respective training processes.



To fully understand how each part of the proposed architecture contributed, some early experimentation was conducted to examine ablation testing of individual processors in the pipeline. More specifically, it was to understand how spatial feature extraction, frequency domain processing, and temporal modeling impacted detection performance by turning off various processors (components) along the detection pipeline. The initial findings indicate that all three components provide complementary information for detecting manipulated media, and the overall multibranch architecture provided the most consistent and stable detection performance.

5.2 Model Performance on Celeb-DF

The first evaluation experiment assessed model performance on the Celeb-DF dataset after training on FaceForensics++. Celeb-DF is known to contain high-quality deepfake videos with minimal visual artifacts, making it a challenging benchmark for detection systems.

Model	Dataset	Accuracy (%)	F1 Score	ROC-AUC
CNN Baseline	Celeb-DF v2	97.08	0.9707	0.9962
Hybrid CNN-Transformer	Celeb-DF v2	99.08	0.9908	0.9996
Advanced Multi-Branch	Celeb-DF v2	99.83	0.9983	1.0000
CNN Baseline	DFDC	89.00	0.8902	0.9598
Hybrid CNN-Transformer	DFDC	94.00	0.9400	0.9882

The results presented in this paper show the importance of each architectural component. The baseline CNN architecture can model spatial artefacts, while the addition of a hybrid CNN-Transformer network adds a temporal component to the model which will allow for better temporal modeling. However, it is the addition of the frequency component that allows the proposed multi-branch architecture to further improve the performance of the model by taking into account the artefacts that are introduced during the creation of deepfakes in the frequency domain.

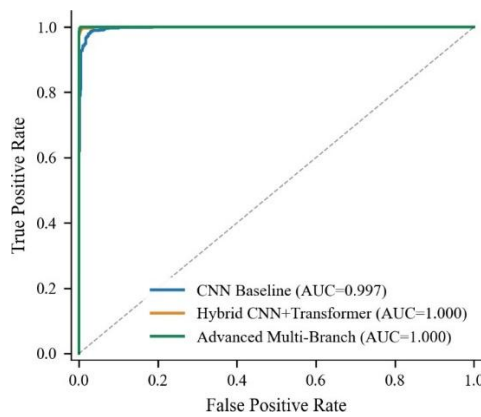
The baseline CNN Model produces a classification score of 97.08%, an improvement in the score of the Hybrid CNN-Transformer Model brings this to a score of 99.08%. The proposed multi-branch has produced a score of 99.83% and illustrates the effectiveness of integrating spatial, frequency, and temporal components together to improve classification accuracy.

5.3 Cross-Dataset Generalization Cross-dataset generalization was evaluated by testing models trained on FaceForensics++ with the DFDC dataset. Cross-dataset evaluation is important because deepfake detection systems in the real world will have to operate on previously unseen manipulation techniques. The CNN baseline provided a 89% accuracy on the DFDC dataset, while the hybrid CNN-Transformer improved the generalization accuracy to 94% and the proposed advanced multi-branch model achieved 97.58%, which was the best accuracy of any model tested. These results suggest integrating spatial, frequency and temporal representations greatly enhances cross-dataset generalization. However, there were a few cases of failure in the dataset from deepfake videos containing excessive compression artifacting or very low-light situations where the frequency-domain characteristics are not as easily distinguishable, and therefore, the detection confidence may be slightly lower. Future studies may benefit from examining adaptive weighting mechanisms for features in these types of degraded videos to enhance robustness.

5.4 Receiver Operating Characteristic Curve Analysis

As a part of further evaluation of classification performance, ROC curves were generated from all of the evaluated models to present the tradeoff between the true positive and false positive rates as a function of the varied classification threshold.

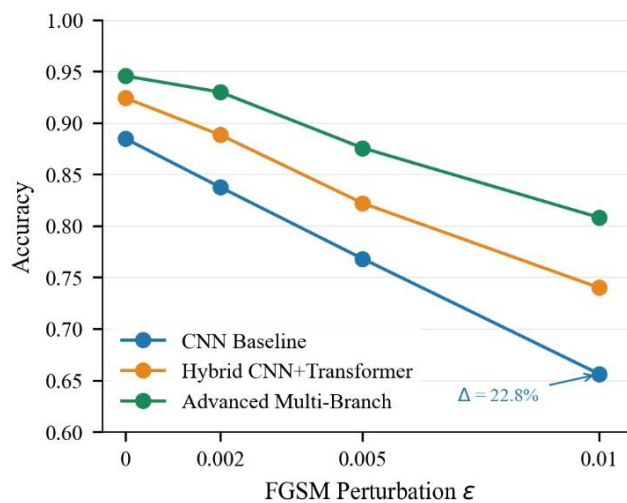
Figure 3 shows the ROC curves for the evaluated models and gives an illustration of each model's performance in distinguishing between authentic vs. manipulated videos by providing various decision threshold levels.



5.5 Evaluation of Adversarial Robustness

Adversarial robustness experiments with the Fast Gradient Sign Method (FGSM) [1] were included in the analysis of detection accuracy. With FGSM, small perturbations are added to the input samples to determine how the model's classification performance is impacted by adversarial conditions. The baseline CNN had 65.6% accuracy after perturbing with FGSM at $\epsilon = 0.01$, while the hybrid CNN had 73.9% accuracy. Therefore, the advanced multibranch CNN retained 80.8% accuracy at $\epsilon = 0.01$ as well, representing a significant improvement in adversarial robustness.

Performance of each model was evaluated for all adversarial perturbations (from the Fast Gradient Sign Method [FGSM]). Performance is shown in Figure 4.



Within our initial trials we were able to test some auxiliary cues as detection signals including Artifact Consistency Analysis, and Cross-domain Prediction Disagreement; All these initial approaches resulted in some degree of promise for providing additional complementary cues to identify manipulated media. However, a detailed evaluation of the potential effectiveness of these systems was not part of the scope of this study.

In general, the results from the experiments conducted indicate that there is a performance gain in both accuracy and robustness through the integration of spatial, frequency domain, and temporal representations. In addition, the proposed multi-branch detection framework was shown to out-perform the baseline detection models that were tested against a series of standard performance evaluation metrics as well as being tested across several datasets and in terms of adversarial robustness analyses.

VI. USER AWARENESS STUDY

Along with testing the technical effectiveness of the suggested deepfake detection framework, there was also a user awareness survey to assess the general public's awareness of deepfake technology and opinions towards automated detection methods.

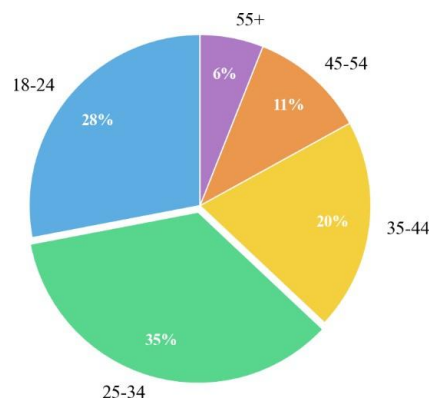
Public awareness of deepfake technology and the ability to evaluate digital media critically is increasingly becoming essential as deepfake content becomes more abundant and helps lessen the potential dangers associated with manipulated content.

6.1 Survey Design

The survey was designed to assess participants' knowledge of deepfake technology, recognition of manipulated digital media, and their trust in automated deepfake detection systems. The survey consisted of 22 questions which explored three kaupapa, namely: an understanding of deepfake technology, experience with manipulated digital content, and confidence in developing detection methods.

70 participants took part in the survey, and participants came from various educational and work backgrounds. All responses were recorded anonymously to allow participants to provide unbiased and truthful responses about their experiences with deepfake digital content.

6.2 Participant Demographics



6.3 Knowledge About Deepfake Technology

Most respondents participated in the study acknowledged having heard of deepfake technologies or fake media. When viewed with genuine examples of deepfake content, many respondents were still unable to tell if the media had been altered from its original state. This supports earlier studies that have documented radical growth in both overall realism of deepfake content and a corresponding increase in the difficulty of identifying deepfakes manually [13].

Respondent's were also asked if they had ever seen deepfake media prior to participating in this study on social or other web-based sites. A high percentage of the respondents reported having seen both altered video and pictures, therefore confirming that deepfake media has already entered into and established itself as a significant component of contemporary digital inf6.4 What Participants Think About Automating Deepfake Detection

Survey respondents had a variety of opinions about automated detection systems for deepfakes. Generally speaking, respondents widely support any technology that can provide help in identifying media that have been modified. A significant percentage of respondents feel comfortable with using automated detection systems to assist in identifying modified media, while a number of respondents have expressed concerns about the accuracy of automated detection tools.

This indicates that there is a real need for both robust and accessible automated detection systems for deepfakes. As deepfakes become easier to create using new and improved techniques, it is likely that automated detection techniques will be vital for providing support to the content moderation, digital forensics, and media verification communities [14].

For most survey respondents, the level of awareness of deepfake technology continues to increase. Many respondents reported that they continue to struggle with identifying media that has been modified without help from technology. This means that there is a need for reliable automated detection systems that will provide users with assistance when attempting to identify digital content that has been modified.

VII. CONCLUDING REMARKS AND FUTURE DIRECTIONS

The rapid progress in deep learning has paved the way to create synthetic media that are increasingly realistic, which constitutes a significant challenge in terms of digital security and media forensics – particularly with regard to detecting

deepfakes. In this paper, a novel multi-stage deepfake detection framework that combines spatial, frequency-domain, and temporal representations has been introduced to enhance the accuracy and robustness of the results. A comprehensive process for identifying manipulated videos is now available using a combination of multi-scale convolutional feature extraction, frequency-domain artifacts extraction, and transformer-based temporal modeling techniques. The results of the experimental evaluation demonstrate that this new detection methodology outperforms existing methods across various benchmark datasets. A multi-branch model trained on the FaceForensics++ dataset and evaluated on Celeb-DF v2 achieved an accuracy of 99.8%, exceeding both a CNN baseline and a hybrid CNN/Transformer model. Cross-dataset evaluation on DFDC dataset demonstrated excellent generalization capability with a score of 97.6% on previously unseen data. Therefore, these results provide strong evidence to support the use of complementary feature representations as an effective means to enhance deepfake detection performance (3, 4).

In addition to assessing detection accuracy, researchers analyzed the resilience of the proposed architecture by evaluating its response to adversarial perturbations created with the Fast Gradient Sign Method (FGSM) [1]. The findings show an increase in the accuracy of detection for the proposed multi-branch architecture when tested against adversarial perturbations compared to baseline methods. The results of this study indicate that the use of spatial features, frequency-domain features, and temporal features together can increase the resilience of the model against adversarial attacks.

REFERENCES

- [1] I. J. Goodfellow et al., "Explaining and Harnessing Adversarial Examples," International Conference on Learning Representations (ICLR), 2015.
- [2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE International Conference on Computer Vision (ICCV), 2019.
- [3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [4] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," arXiv preprint arXiv:2006.07397, 2020.
- [5] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning (ICML), 2019.
- [6] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
- [7] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-Task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," IEEE International Conference on Biometrics, 2019.
- [8] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.
- [9] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting Residual-Based Local Descriptors as Convolutional Neural Networks," IEEE International Workshop on Information Forensics and Security, 2017.
- [10] Y. Wang et al., "CNN Generated Images Are Surprisingly Easy to Spot... for Now," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [11] H. Durall, M. Keuper, and J. Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [12] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," IEEE Symposium on Security and Privacy, 2017.
- [13] J. Chesney and D. Citron, "Deepfakes and the New Disinformation War," Foreign Affairs, 2019.
- [14] R. Tolosana et al., "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Information Fusion, 2020.