ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

Deepfake Classification Using an Attention-Based Multi-Domain Transformer Model

Lalam Sravani[1]; Ch. Pardhiv Kumar[2]; K. Leela Pramod Kumar[3]; A. Venkatesh[4]; B. Nithin[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India.
[2]Student, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India
[3]Student, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India
[4]Student, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India
[5]Student, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India

Publication Date: 2026/05/30

Abstract: In emerging digital media, deepfake videos have already been shown to be an extremely dangerous source of threats to trust, since such technology can create manipulate digital content in videos and images which looks highly realistic. There is a possibility of employing synthetic videos that leads to malicious behavior, which is danger and fraud, so there is a need to introduce robust methods for detecting deepfakes. It should be noted that conventional approaches t detecting deepfakes tend to employ one form of CNN (Convolutional Neural Networks), for instance ResNet and EfficientNet, however, in practice, such algorithms may not be applicable due to large number of methods used for manipulating videos and images. In this study, we propose to employ a hybrid approach based on utilizing deep neural network for hybrid feature extraction to build a better detection system. Specifically, our proposed method involves analyzing both spatial and temporal features by using Swin transformer and Temporal transformer and we also include Frequency transformer.

Keywords: *Deepfake Detection, Deep Learning, Swin Transformer, Temporal Transformer, Frequency Transformer, Feature Fusion, Video Classification.*

How to Cite: Lalam Sravani; Ch. Pardhiv Kumar; K. Leela Pramod Kumar; A. Venkatesh; B. Nithin (2026) Deepfake Classification Using an Attention-Based Multi-Domain Transformer Model . *International Journal of Innovative Science and Research Technology*, 11(4), 5053-5060. https://doi.org/10.38124/ijisrt/26apr2013

I. INTRODUCTION

Development in deepfake technology have been very swift over the recent past that it has become possible for deepfake videos to edit faces along with facial expressions. There are many number of factors that need to be deal with, because deepfake technology have been used in many ways such as to entertain by including deception, misrepresentation and impersonation among others.

Using Generative Adversarial Networks (GAN), Many deepfake videos are generated and to process the generated facial features can be manipulated and each frame can be processed independently by maintaining the smooth flow of movements within the video. Even while being quite invisible to the naked human eye, the alterations in the way certain region move like blinking eyes. The majority of the techniques for deepfake detection employ CNN which is quiet efficient approach but there are certain limitations i.e. Motion signs are not suitable and as well, as they fail to take into account the correlation between frames. The mentioned challenges can be solved by using few approaches that are being developed in order to classify in a better manner and to enhance the performance of those approaches. To recognize the complex interactions that occur during deepfakes generation, models like Swin, temporal and frequency can used. The video sequences have been used in this work to detect deepfakes in those videos by using deep learning models and to detect various types of features through processing them and to extract various artifacts within frequency domain, also to gain fine-spatial information by using the Swin transformer, utilizes a transformation layer to capture temporal information between frames and also to adopts the frequency transformer technique.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

II. LITERATURE REVIEW

In recent works, classical image processing techniques and CNN, machine learning and deep learning models have been used by including the image processing techniques in order to detect deepfake videos automatically that measures the model performance using evaluation metrics which varies trade-offs in accuracy, complexity, and applicability.

Soudy et al. (2024) [1] created a combination of CNNs and VIT in order to classify the fake and real videos. The data pre-processing like like face extraction and other data augmentation techniques have been used to process the DFDC and FaceForensics++ datasets and thus achieved 94.7% accuracy by these combinational models. The limitation within model is high complexity that makes the system impractical for real-time applications.

Gong et al. (2024) [2] created an approach based on Swin transformer design to make the model to learn spatial features. When tested using datasets such FaceForensics++, Celeb-DF and DFDC, it reached accuracy close to 96%. However, the model has a limitation that is unable to find the temporal and frequency domain related features.

Gao et al. (2024) [3], Based on DCT and wavelet transformer methods, the high-frequency artifact detection framework was introduced and trained on FaceForensics++ and DFDC datasets, it achieves close to 94.3% accuracy. However, it comes with challenges of increasing computational requirements.

Luan et al. (2024) [4] proposed a method for detecting the deepfake videos which works by using Frequency -spatial transformer where it gave 92.8% accuracy by testing it on Celeb-DF and FaceForensics++ datasets. The major challenge with this approach is of instability while operating with different types of input.

Hasannath et al. (2025) [5] offered a strategy for deepfake detection with frequency analysis being prioritized. The study relies in a unique frequency enhanced self-blended images (SBI) technique that is capable of producing realistic signs of manipulations. To achieve an outstanding accuracy level of 94.9% by using Celeb-DF and DFDC datasets to train the models and to identify the distinctive feature by addressing generalized capability of detecting manipulated videos previously that are unseen by the model. Through the usage of artificial augmentation of synthetic data, the technique includes training of the algorithm on reality.

Yadav et al. (2025) [6] was proposed an hybrid approach for solving the problem for deepfake detection consists in the combination of two approaches. Thus, the spatial part involves the utilization of a CNN (e.g., Xception, EfficientNet), designed for the detection of the abnormalities in appearance of a person, such as skin texture deformation, problems with the lighting. This includes the FFT-based Frequency-Domain Stream, also allows the decomposition of the input video stream into frequencies in order to detect high-frequency features that are invisible to the human eye and associated with deepfakes. For analyzing the relationships between frames and to spot the temporal inconsistencies such as abnormal blinking, excessive movements can be identified by having video element to use recurrent neural networks (RNN), BiLSTM or Temporal Transformer. Finally, it is concluded from the results, that excellent results are demonstrated in terms of accuracy in deepfake detection, reaching 93.8% in two well-known face forgery databases (FaceForensics++, DFDC). However, this approach fails to identify low-quality and compressed images due to the destruction of frequency information.

By reviewing the existing works that are mostly relied on analyzing the features manually to identify several consistent gaps, most systems rely on manually engineered features, require specialized hardware or controlled environments, are not scalable to web or mobile deployment, and do not leverage the representational power of deep neural networks. Our proposed system addresses these limitations by Tri domain transformer ensemble, which eliminates manual feature extraction, works well with small and large datasets, and delivers state-of-the-art accuracy with a lightweight model suitable for real- time deployment.

III. METHODOLOGY

In this modern digital media, to identify deepfake videos is one of the most challenging thing is to enhance generative models that produce high-quality deepfake videos. So, to examining deepfake videos manually will take much time and may lead to errors due to some minor detail's changes in the video content. To identify them manually and to differentiate between real and fake videos that are numerously inconsistent can be managed by spatial, temporal and frequency domains automatically. So, it is needed to propose a deep learning model that will identify those inconsistencies easily. Therefore, we introduced a Tri domain transformer ensemble where it is a combination of Swin, Temporal and frequency transformers. In the beginning of the project, the frames of the videos are extracted and sorted in a fixed format.

Each frame goes through some pre-processing before the training procedure takes place including the processes like resizing, normalization and data augmentation techniques. Swin transformer is employed by the model to capture spatial features because they allow detecting the hierarchy of features, which is important for context analysis. The Swin Transformer employs the shifted windowed attention approach to learn both global and local features at the same time in order to identify visual artifacts that exist in the fake videos. The above advantage facilitates the model to consider temporal dependency during analysis by the use of sliding windows that are used to learn motion and consistency of expressions from one frame to another as well as transitions from frame to frame.

Beside of considering the spatial and temporal features, it is important to consider the frequency domain features also. In this Frequency transformer, the model is able to perform transformation on image data to convert them from space

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

domain to the frequency domain. Deepfake generation involves some sort of artifacts in the frequency domain of the content, the model considers frequency in the convolutional layers. The fusing technique is considered in order to fuse the extracted spatial, temporal and frequency features into feature maps. The feature maps undergo fully connected layers for real and fake categorization.

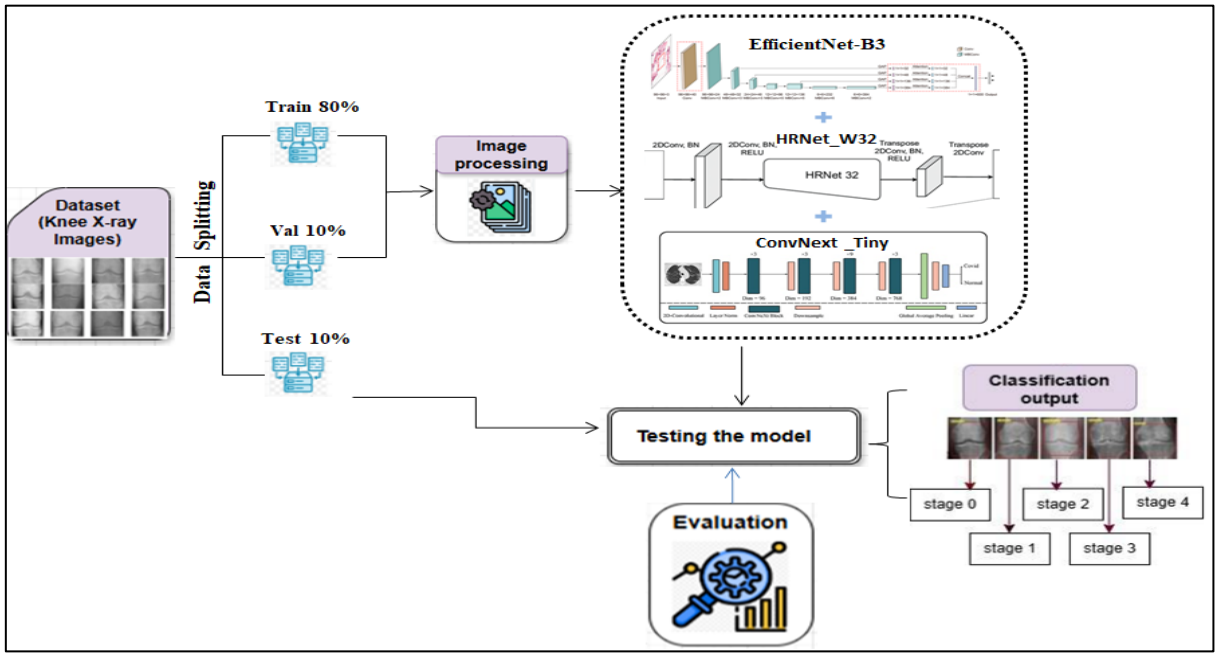

Fig 1 Architecture of the Proposed Model

➢ *Phase-1: Dataset Collection*

In this project, we have used a wide set of videos from various sources on Kaggle in order to ensure both a greater variety and increased reliability. The basis of the whole project is the Deepfake Detection Dataset that contains plenty of videos modified with the help of several methods for creating deepfakes. The classes are divided into fake (0) and real (1). Initially there was an unbalance between those classes since there were more fake videos than real ones. In order to make up for the lack of real samples, we have utilized other reliable resources like the Deepfake Detection Cropped Dataset and FaceForensics++. The final dataset includes 6523 videos where there were of 3455 real and 3068 fake videos.

- *Splitting the Data*

This dataset has been divided into three categories: Training, Validation and Testing, in the ratio of 70%, 15%, 15%. The training category is where the major share of the data lies and is used to train the model to detect patterns that distinguish real and fake videos. In the process of training, the validation category evaluates the model hyperparameters by regulating them to avoid overfitting. The testing category acts as a measure of evaluating the model on unseen data. The real and fake videos have been equally distributed among the three categories to ensure neutrality and reliability.

➢ *Phase-2: Frame Extraction*

Video clips ae made up of a series of continuously captured frames containing temporal information over a period of time. For the effective processing of videos using deep learning techniques, it becomes necessary to break down those videos into individually frames because many computer vision algorithms are image-centric and cannot be used directly for analyzing video content. In this research work, all videos undergo an operation where selected frames are extracted by using uniform sampling. Selection is done by using equal frame intervals from the video sequence. These frames come from different parts of the video, including the beginning, middle, and end sections of the clip. In this frame extraction the frames extracted by every 10th frame of the video and up to maximum of 30 frames of a video.

Each frame is re-sized to a uniform resolution and then transformed into a uniform format that can be used by the model. The sequence of frames is kept intact since this will enable the system to learn the inconsistencies in movement within the video. This process of converting video to an organized series of frames enables the handling of the data without any loss of important information related to time and space.

➢ *Phase-3: Image Processing*

It is essential before introducing the images to the algorithm we wish to use to make sure that all the images have the same dimension, format, and representation. Preprocessing might help in standardizing the images to be used during training. The images need to be standardized such that differences in the quality of images do not hinder the learning process. It is important to remember that there are three different axes for the image, height, width, and channels. The image will be represented by the following three-dimensional axes. H stands for Height, which is the number of pixels along the y-axis, W stands for Width which is the number of pixels along the x-axis, while C represents the channels of the image.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

- *Resizing*

The initial extracted video frames have spatial dimensions denoted as (H, W, C), where H stands for height, W stands for width, and C stands for the colour channels (RGB). As deep learning models need fixed dimensions for optimal operation, the extracted frames undergo rescaling to maintain a constant size of (224, 224, 3).

It is necessary to resize because it guarantees that all images have a common spatial dimension. This helps during batch processing and feature extraction. It further reduces computation by ensuring that all images have a common dimension. Therefore, the network concentrates on extracting features from the images without focusing on their spatial dimensions. Mathematically, resizing can be expressed as:

I^'="Resize" (1)

- *Data Augmentation (Training Phase Only)*

The augmentations that are used during training will help to enhance model generalization that include:

✓ *Random Horizontal Flip:*

The frames will be rotated in such a way that there is an equal chance for each image to be rotated either way. The rotation will make the model learn the characteristics of faces that would not be affected by their orientation in any way. Due to the symmetry of the human face, this measure can be particularly helpful in spotting deepfakes.

I″ = RandomFlip (I ‘rotated) … equation (2)

✓ *Random Rotation:*

To simulate variations in head movement and camera angles, images are randomly rotated by a small angle. This is intended to ensure that the trained model will be capable of handling small discrepancies in angle.

I′rotated = RandomRotate (I′) … equation (3)

By applying these augmentations, the model can easily generalize on unseen dataset. These alterations ensure that the model can effectively detect deepfake content even when faces are not perfectly aligned or oriented.

- *Color Jittering*

To further improvement for robustness of the model, color jittering is applied by randomly changing the brightness, contrast, and saturation of the input frames. This transformation simulates variations caused by different lighting conditions, camera settings, and environmental factors, which are commonly observed in real-world video data.

I^'"="ColorJitter"(I_"rotated”)

Where I_"rotated" is the input image after rotation, and I^'" represents the augmented image after applying color transformations.

Such an addition ensures that the model is not overly dependent on the particular color signs which can vary from one dataset to another. This would ensure that the resulting model works efficiently irrespective of the lighting conditions and other aspects.

- *Tensor Conversion*

In order to train a deep learning model, the image is transformed into tensor pitch (PyTorch) and then restructured into channel-first format:

I′tensor = ToTensor(I″) … equation (4)

Where:

Is the tensor converted image with coordinates (C,H′,W′).

- *Normalization*

For effective training, we make use of standard ImageNet normalization techniques for scaling the inputs of the image pixel values. To maintain numerical stability and convergence, we applied this approach throughout the training phase.

For stable and efficient training, pixel values are normalized using ImageNet mean (μ) and standard deviation (σ).

$$I\_norm = (I\_tensor-\mu)/\sigma$$

Where:

$$\mu= [\mu1, \mu2, \mu3], \sigma= [\sigma1, \sigma2, \sigma3]$$

For the validation and testing datasets we only do the steps to get the data ready. This includes making the images the size changing them into tensors and making sure the numbers are all on the same scale. We do not add any data to these datasets.

I_val=Normalize (ToTensor (Resize (I)))

The preprocessing steps for validation and testing datasets are such as resizing, conversion of tensors and normalization.

➢ *Phase-4: Model Building*

In this phase, an ensemble of deep learning models is built for classifying videos into real and fake categories using multi-domain features. This approach is known as Tri-Domain Transformer Ensemble (TDTE), which consists of multiple transformers that include spatial, temporal, and frequency-based learning. In TDTE, a series of video frames where later those frames are resized into dimensions of 224 × 224 is used as input. These frames are then normalized and transformed to tensor format.

The model uses Swin Transformer as the backbone for capturing spatial information. The architecture enables the model to extract information from both local and global perspectives by using shifted window self-attention

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

mechanism, thereby enabling it to detect hidden details, such as texture discontinuity, incorrect compositing, and unusual face appearance, which are evident in deepfake images.

To find temporal dependencies, a Temporal Transformer Encoder is incorporated. Analysis of the order of the frames is carried out by this module. The relations among them are captured using self-attention over time. The analysis of temporal order helps detect inconsistencies in motion, facial movements, and transition between frames, which are indicative of video manipulation. Through the consideration of all temporal data, overall video understanding is improved.

In addition to spatial and temporal information we also considering frequency cues, we introduce a new branch to find hidden artifacts. Using FFT to transform input frames into the frequency domain, frequency features are extracted by passing those frames through a series of convolutional layers. Through such a process, it is possible to identify any synthesis artifacts present in the spatial domain.

Then, the extracted features from all three domains are combined into a single feature vector to incorporate all the spatial, temporal, and frequency aspects. The resulting combined feature is passed through fully-connected layers followed by a SoftMax classifier for determining whether the input video clip is authentic or tampered.

Through the use of transformers for handling spatial aspects, sequence learning for the temporal domain, and frequency components, the proposed TDTE framework provides reliable results for deepfake detection.

➢ *Phase-5: Model Evaluation*

The suggested TDTE algorithm is analyzed based on the capacity of this system to authentic videos and fake (0) videos real (1). To make sure that the results of this analysis are reliable and unbiased, this algorithm is evaluated using an unseen testing data set. Evaluations are made using some well-known performance parameters like accuracy, precision, recall, and F1-score. Accuracy means the number of predictions that have been correct, and precision indicates the level of perfection in detecting fake videos. Recall means recognizing all the instances where fake videos can be detected, while the F1-score is a combination of precision and recall.

One more method for evaluating the model involves confusion matrix in which there will be shown visually how predictions of the model were – true positives, true negatives, false positives, and false negatives.

IV. RESULTS

The deepfake detection problem is classified upon the proposed model Tri domain Transformer Ensemble (TDTE) which integrates spatial, temporal and frequency features are evaluated on the deepfake detection validation and testing datasets to classify whether the video is real (1) or fake (0).

The individual components of the model such as swin transformer, temporal transformer and frequency transformer tend to learn both local and global features of the frames. This TDTE model achieves validation accuracy 95.4% and testing accuracy 94.7%, indicating strong generalization capability. They are measured using typical metrics for accuracy, recall and F1-score found in a classification report. We use a confusion matrix to visualize how often our model is wrong vs. predicting the correct labels. As you can see from the experiment, the model helps reduce false positives and negatives. Additionally, graphs showing our accuracy and loss during training show steady performance with no overfitting. Employing various feature spaces allows for an anomaly to be detected using all kinds of spatial, temporal and spatial outliers.

➢ *Training and Validation Performance*

The proposed model was trained for 10 epochs with a batch size of 4 on the constructed deepfake video dataset consisting of real and fake samples. During training phase, the model demonstrated with training accuracy reaching of 95-96%. The training loss decreased steadily across epochs, indicating effective learning of spatial, temporal and frequency-based features without significant overfitting.

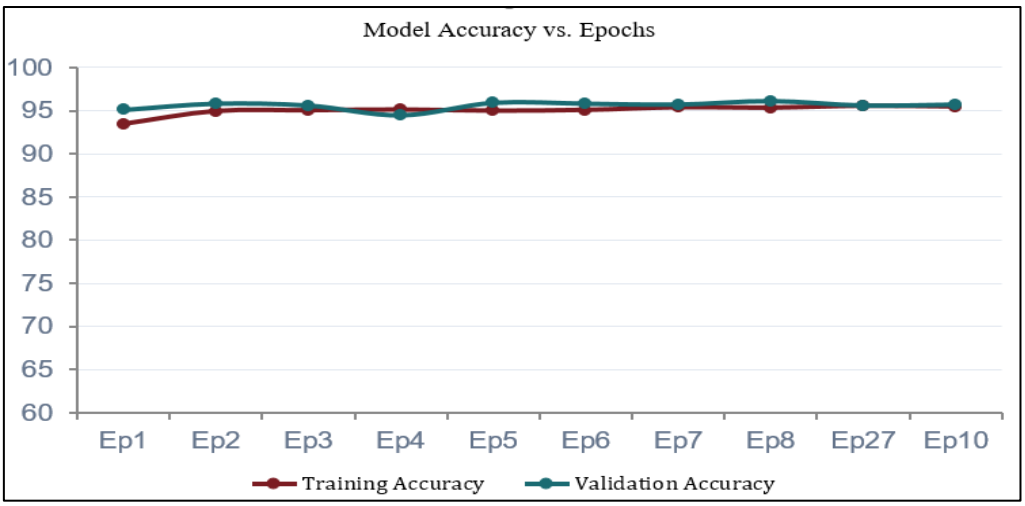

Fig 2 Model Accyuracy vs. Epochs

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

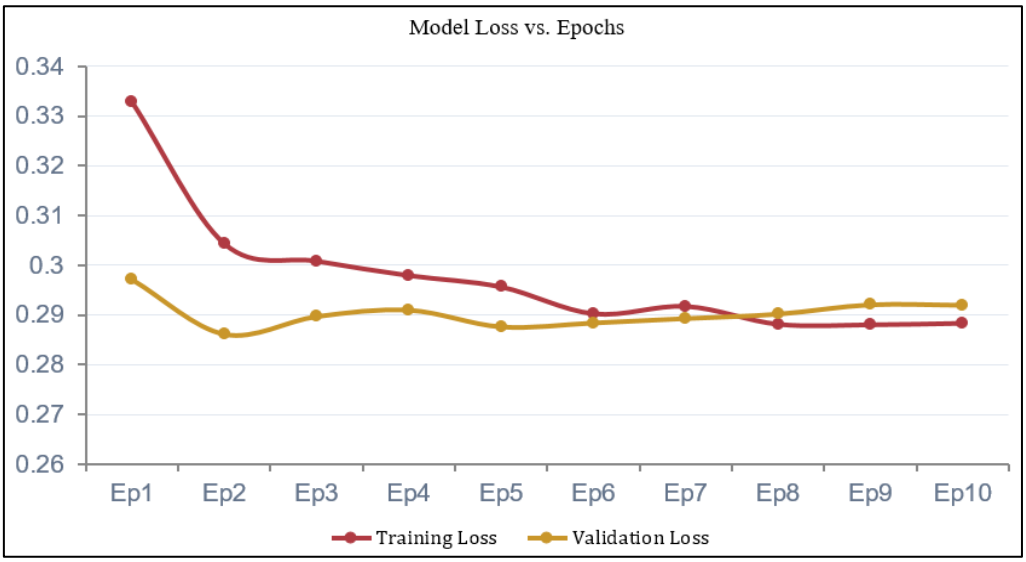

Fig 3 Model Loss vs. Epochs

Classification Report

	precision	recall	f1-score	support
Real	1.00	0.91	0.95	510
Fake	0.91	1.00	0.96	469
accuracy			0.96	979
macro avg	0.96	0.96	0.96	979
weighted avg	0.96	0.96	0.96	979

Fig 4 Classification Report

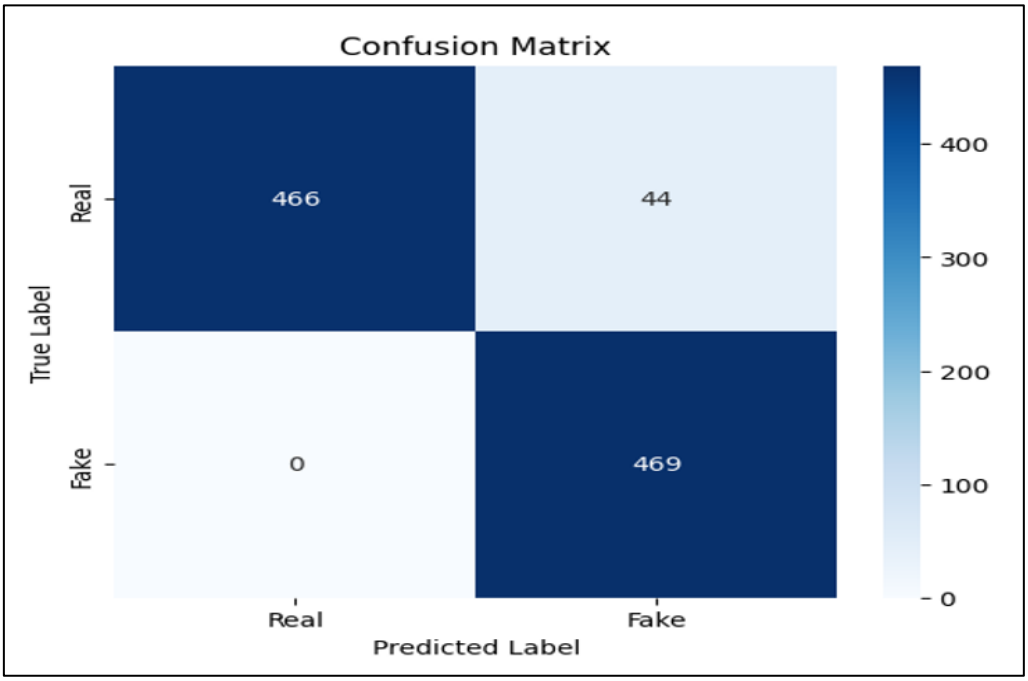

Fig 5 Confusion Matrix

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

V. RESULTS DISCUSSION

The proposed TDTE model combined spatial, temporal and frequency domain features, resulting improved deepfake detection performance compared to single domain models. The model achieved a test accuracy of 95.5%, demonstrating strong capability on classifying the real and fake videos. The integrated model learning approach enhances the model ability to capture the spatial and temporal artifacts and hidden frequency patterns in deepfake detection.

The overall classification performance is supported by balanced evaluation metrics, with precision, recall, and F1-score values close to 0.94, indicating consistent performance across both real and fake classes. The model maintains high recall for fake videos, which is critical for minimizing false negatives in deepfake detection systems.

From an individual perspective, the Swin Transformer that extracts spatial features performs well due to its ability to capture minute visual features. The temporal transformer improves the performance since it captures the dependency between frames and the motion inconsistencies. Frequency branch contributes to improved performance since it is able to identify spectral artifacts which cannot be detected from the spatial domain. Combining the three elements' results in a model that outperforms using individual feature domain models.

The training and validation accuracy and loss curves indicate smooth convergence without significant overfitting. The small gap between training and validation metrics confirms that the model generalized well on unseen data.

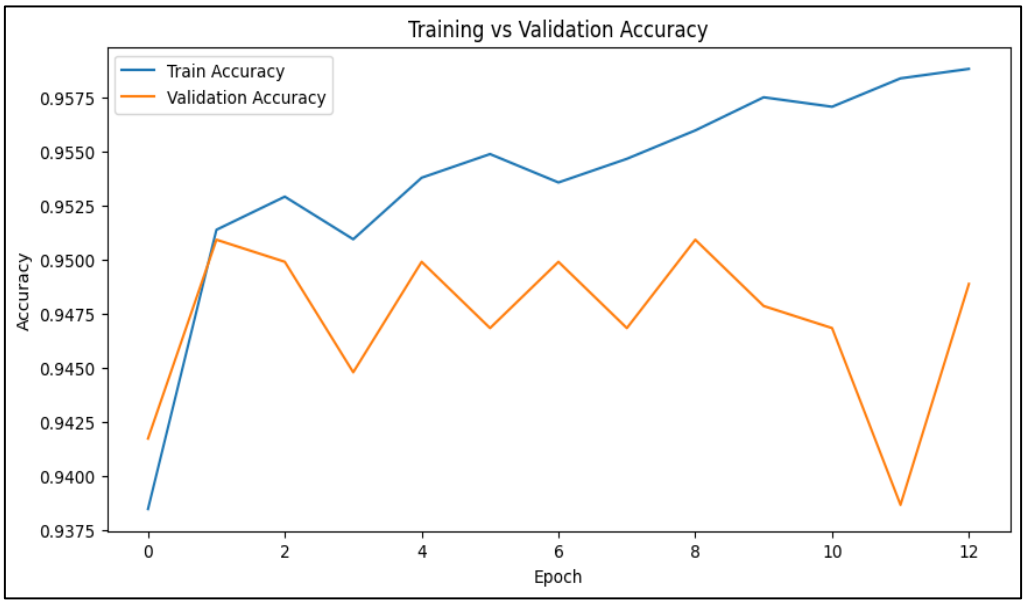

Fig 6 Training vs Validation Accuracy

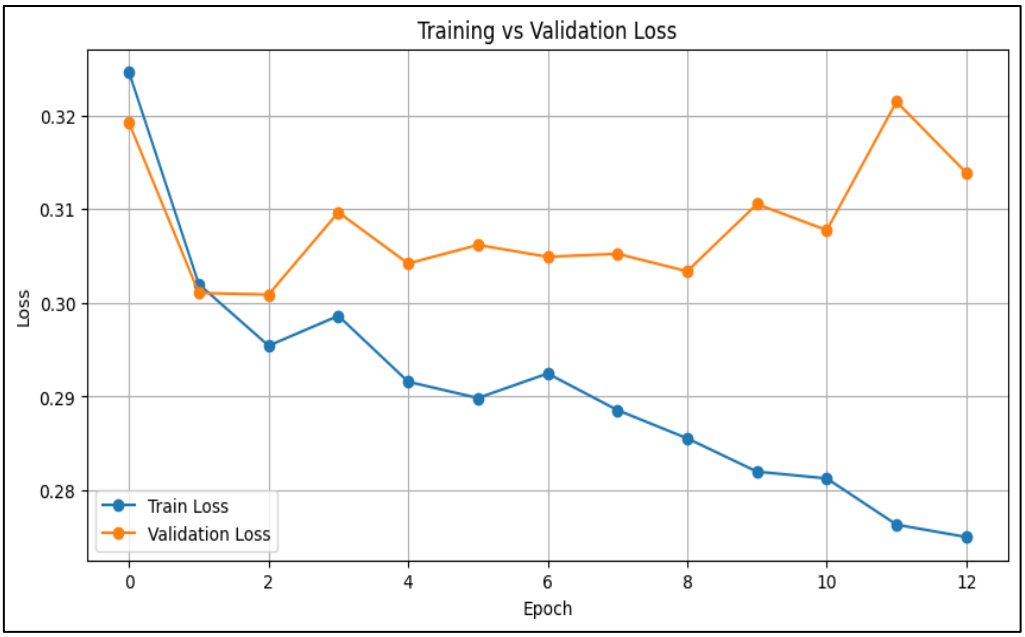

Fig 7 Training vs Validation Loss

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2013

VI. CONCLUSION

By using a combination of spatial, temporal, and frequency-domain features into one deep learning architecture, the method used in detecting deepfakes becomes efficient and reliable. This is due to the use of the strength of the Spatial features through the use of the Swin transformer, the temporal transformer to capture sequences in videos, and FFT in determining hidden frequencies that make deepfakes unique. Together, all the domains used in this technique make the TDTE model become a powerful deep learning model that classifies deepfakes.

From the study, it is evident that there is an importance of utilizing multi-domains in order to solve some of the problems associated with detecting deepfakes. In addition, the model performed very well when tested with real datasets thus making it efficient in detecting deepfakes.

For future research, attention will be focused on the use of larger datasets with different variations to make the deep neural network model efficient. Other improvements will include making it easier to deploy to production by reducing its size and also introducing explainability into the model.

REFERENCES

[1]. Soudy, Ahmed Hatem, et al. "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks." Neural Computing and Applications, vol. 36, no. 31, 8 Aug. 2024, pp. 19759-775, https://doi.org/10.1007/s00521-024-10255-y.

[2]. Gong, Liang, Xue Li, and P. H. J. Chong. "Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector." Electronics, vol. 13, no. 15, Aug. 2024, p. 3045, https://doi.org/10.3390/electronics13153045.

[3]. Gao, Jie, et al. "DeepFake Detection Based on High-Frequency Enhancement Network for Highly Compressed Content." Expert Systems with Applications, vol. 249, part A, Aug. 2024, p. 123732, https://doi.org/10.1016/j.eswa.2024.123732.

[4]. Luan, Tao, Guoqing Liang, and Pengfei Peng. "Interpretable DeepFake Detection Based on Frequency Spatial Transformer." International Journal of Emerging Technologies and Advanced Applications, vol. 1, Mar. 2024, pp. 19-25, https://doi.org/10.62677/IJETAA.2402108.

[5]. Sunil, Reshma, et al. "Exploring Autonomous Methods for Deepfake Detection: A Detailed Survey on Techniques and Evaluation." Heliyon, vol. 11, no. 3, Feb. 2025, p. e42273, https://doi.org/10.1016/j.heliyon.2025.e42273.

[6]. Li, Yi, et al. "A Generalizable Deepfake Detection Method Based on Local Spatial-Frequency Feature Fusion." Proceedings of the 2026 International Conference, Jan. 2026, pp. 9-15, https://doi.org/10.1145/3779153.3779155.

[7]. Hasanaath, Ahmed, et al. "FSBI: Deepfake Detection with Frequency Enhanced Self-Blended Images." Image and Vision Computing, vol. 154, Feb. 2025, p. 105418, https://doi.org/10.1016/j.imavis.2025.105418.

[8]. Yadav, Uma, et al. "A Hybrid Approach for Robust Deep Fake Image Detection Using Spatial and Frequency Domain Features." Engineering, Technology & Applied Science Research, vol. 15, no. 3, June 2025, pp. 22786-791, https://doi.org/10.48084/etasr.10458.

[9]. Chorage, S. S., et al. "Deepfake Detection Using Deep Learning." International Research Journal on Advanced Engineering Hub, vol. 3, no. 8, Aug. 2025, pp. 3427-31, https://doi.org/10.47392/IRJAEH.2025.0502.

[10]. Iliev, Alexander I. "Discovery of Deepfakes in Art." Digital Presentation and Preservation of Cultural and Scientific Heritage, vol. 15, Sept. 2025, pp. 55–64, https://doi.org/10.55630/dipp.2025.15.5.

[11]. Spatiotemporal Deepfake Video Detection: A Hybrid CNN-Transformer Approach with Frequency Analysis." 2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI), IEEE, Sept. 2025, https://doi.org/10.1109/IRI61234.2025.00012.

[12]. Usman, Muhammad, et al. "Lightweight and Hybrid Transformer-Based Solution for Quick and Reliable Deepfake Detection." Frontiers in Big Data, vol. 8, Mar. 2025, https://doi.org/10.3389/fdata.2025.1521653.

[13]. "HTMDF-DD: Hybrid Triple Modality Based Spatial–Temporal Features Early Fusion for Deepfake Detection." ResearchGate, Feb. 2026, https://www.researchgate.net/publication/400322321.