2024 28th International Conference on Computer and Information Technology (ICCIT)
20-22 December 2024, Cox's Bazar, Bangladesh

TransCNN: A Hybrid CNN-Transformer Synergy for Reliable Deepfake Forensics

Md. Sabbir Hossen and Md. Saiduzzaman
Dept. of Computer Science & Engineering, Bangladesh University, Dhaka-1207, Bangladesh
Email: sabbir.hossen@bu.edu.bd

***Abstract*—The rapid advancement of deepfake technology has raised serious concerns regarding digital authenticity, misinformation, and media trust. By enabling the creation of hyper-realistic manipulated media, deepfakes pose significant risks to security, privacy, and the integrity of digital communication. Social platforms are increasingly flooded with such AI-generated content, which can deceive human perception, spread misinformation, and undermine the credibility of multimedia information. Therefore, early detection of deepfakes is crucial. In this research, we focus on developing a robust detection framework to distinguish between genuine and artificially generated deepfake face images. For this purpose, a dataset of 12.2k authentic images was collected, and an additional 12.2k deepfake images were generated using a generative adversarial network (GAN), ensuring a balanced dataset for real versus fake image classification. We present a custom CNN-transformer hybrid framework, TransCNN, for deepfake detection. The proposed framework leverages convolutional layers for spatial feature extraction and transformer-based attention mechanisms to capture contextual dependencies within facial patterns. The experimental evaluation demonstrates that the proposed CNN-based Transformer, TransCNN model, achieves an exceptional accuracy of 98.91%. This study underscores the importance of hybrid deep learning approaches in addressing the growing challenge of deepfake detection and highlights their potential for safeguarding digital authenticity, particularly within the domain of facial image verification.**

***Index Terms*—Deepfake Detection, Convolutional Neural Network, Transformer Model, Generative Adversarial Network.**

I. INTRODUCTION

In recent years, advances in artificial intelligence have given rise to deepfake technology, a powerful yet controversial innovation capable of producing hyper-realistic manipulated images and videos that closely mimic authentic media. Leveraging generative adversarial networks (GANs) and other advanced neural architectures, deepfakes can convincingly alter facial expressions, synthesize entirely new identities, and even generate fabricated speech [1]. While this technology highlights the creative potential of generative AI in fields such as entertainment, education, and digital content creation, its misuse poses severe threats to digital security, personal privacy, and public trust. Malicious applications of deepfakes range from spreading misinformation and political propaganda to identity theft, financial fraud, and cybercrime [2]. The increasing sophistication and accessibility of these tools have led to a surge of AI-generated media across online platforms, creating an urgent demand for reliable detection systems capable of preserving the integrity and authenticity of digital information. The challenge of deepfake detection lies in the subtle visual and contextual cues embedded in generated media, which can be imperceptible to the human eye [3]. Traditional computer vision methods often fail to generalize against such highly realistic manipulations. As a result, machine learning and deep learning have become critical tools for addressing this pressing challenge [4], [5]. CNN-based models are effective in extracting spatial features from facial structures, while transformer-based methods excel at capturing contextual dependencies. Combining these two approaches provides a promising avenue for building robust and generalizable detection systems [6]. In this research, a custom CNN-Transformer model, termed TransCNN, was developed by integrating convolutional and transformer architectures to effectively distinguish between real and deepfake images. `The main contributions of this study` are outlined as follows. A custom hybrid CNN-Transformer framework was developed for deepfake image detection, leveraging the complementary strengths of CNNs in local feature extraction and Transformers in global contextual modeling. A dataset of deepfake images was created using a Generative Adversarial Network (GAN) to provide realistic synthetic samples for training and evaluation. Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to generate visual explanations and enhance the interpretability of the model's predictions. Finally, the effectiveness of the proposed approach was validated through a comprehensive performance evaluation using standard classification metrics, including accuracy, precision, recall, and F1-score. `The remaining sections of this study` are organized as follows. Section II provides a detailed review of the existing literature on deepfake detection, highlighting the current challenges. Section III describes the dataset preparation and methodology. Section IV presents the experimental setup and implementation. Section V presents the evaluation results and discusses the strengths and limitations of the proposed model. Finally, Section VI concludes with a summary of findings and future research directions.

II. LITERATURE REVIEW

The rapid progress of generative models drives the emergence of deepfakes. This growing phenomenon has prompted significant research interest, as the malicious use of deepfakes can lead to severe implications ranging from misinformation and political manipulation to privacy violations and cybercrime. As the realism of generated content improves, traditional detection methods are often rendered less effective, highlighting the need for more robust and adaptive solutions. To better understand the evolution of this domain, a review of existing studies was conducted, focusing on the techniques, challenges, and research gaps associated with deepfake detection. For instance, *S. Sohail*

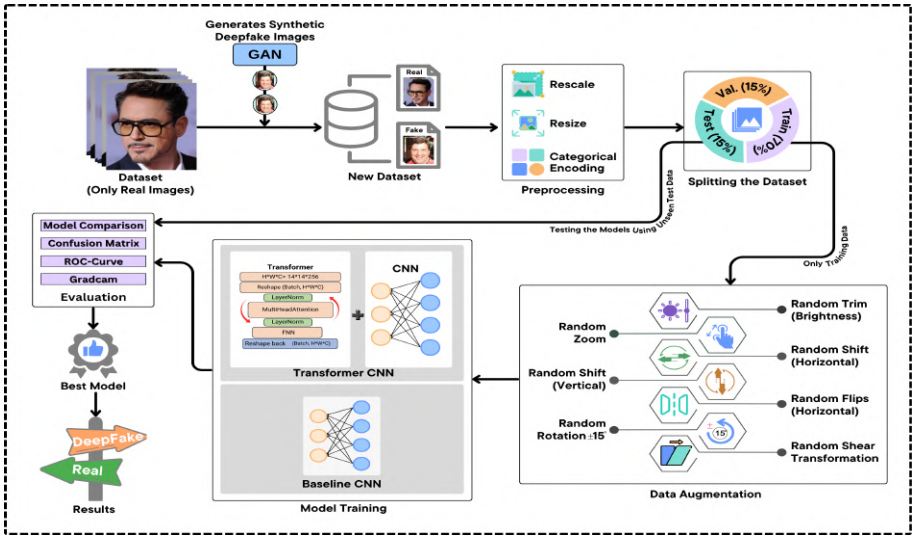

Fig. 1. The Proposed workflow diagram for Deepfake Detection

et al. [7] explored deepfake detection using CNNs, RNNs, and hybrid models such as CNN-LSTM, CNN-GRU, and TCNs to capture both spatial and temporal features. They incorporated GAN-based data augmentation and a fusion of artifact inspection with facial landmark detection, achieving over 99% accuracy across datasets, though challenges remain with compressed formats, noise, and generalization. *F. Zafar et al.* [8] introduced a lightweight deepfake detection framework combining EfficientNetB0 with Temporal CNNs, supported by MTCNN-based face alignment and augmentation strategies like CutMix, MixUp, and RandomErasing. Their model leveraged FPN for multi-scale feature fusion, attaining 92.45% accuracy on FFIW-10K with only 0.45 GFLOPs, balancing accuracy with computational efficiency for real-world deployment. *I. Ambreen et al.* [9] proposed a hybrid detection model integrating Vision Transformers with CNNs to capture fine-grained spatial inconsistencies in facial patterns. Evaluations on 76,161 images achieved 99% accuracy, precision, recall, and F1-score, demonstrating the superiority of transformer-based architectures over conventional CNNs and underscoring their potential for reliable and robust deepfake detection in diverse scenarios. *A. Kumar S. et al.* [10] presented a hybrid approach that integrates YOLOv8 with RNNs for deepfake detection, where YOLOv8 extracts spatial features while RNNs capture temporal dependencies and subtle inconsistencies across frames. Their method efficiently identifies manipulated content in images, providing an effective framework for detecting the misuse of deepfake technologies in practical applications. *D. Awasthi et al.* [11] developed a multi-resolution deepfake detection algorithm incorporating Viola-Jones face detection with RDWT, MSVD, and DCT for frequency feature extraction. ANFIS-based optimization improved robustness, while deep CNNs (SqueezeNet, ResNet50, EfficientNet, InceptionV3) enhanced performance. Their method improved imperceptibility by 52.14% and robustness by 7.51%, offering a strong solution for secure image authentication. *A. Jaiswal et al.* [12] proposed an efficient CNN–VGG16 hybrid model for deepfake detection, trained on benchmark datasets. Their framework achieved 95% accuracy and 94% precision, outperforming many existing approaches. The model demonstrated strong generalization across manipulated samples, proving the effectiveness of integrating VGG16 features with CNN-based architectures for detecting forged visual content. *R. Sharma et al.* [13] investigated CNNs and advanced architectures, including ResNet50 and XceptionNet, for deepfake detection using real and synthetic datasets. XceptionNet achieved the highest performance with 84% accuracy, while preprocessing improved feature quality and reduced noise. Their work highlighted trade-offs across different architectures and emphasized the importance of preprocessing for achieving robust detection outcomes. *K. Magoo et al.* [14] introduced a CNN-based detection framework addressing both deepfake images and forged signatures. Their system was trained on diverse authentic and manipulated samples, achieving superior accuracy and adaptability compared to conventional techniques. This robust framework contributes to digital forensics and cybersecurity, offering practical strategies for content authentication in increasingly complex manipulation environments. Despite significant progress in deepfake detection using CNNs, RNNs, Vision Transformers, and hybrid architectures, current methods face limitations. Many models achieve high accuracy only on specific small datasets, lacking robustness against compression, noise, or unseen manipulation techniques. Additionally, challenges remain in balancing accuracy with computational efficiency, improving generalization across domains, and integrating explainability to ensure forensic reliability and applicability in real-world scenarios.

III. MATERIALS AND METHODS

This section presents the research methodology of this paper. The process includes dataset description, deepfake generation, image preprocessing, data augmentation, and building the hybrid CNN-Transformer model. The proposed workflow diagram is illustrated in Fig. 1.

A. Dataset Collection and Preparation

In this research, the celebrity face image dataset was utilized, which was collected from a public source [15]. The

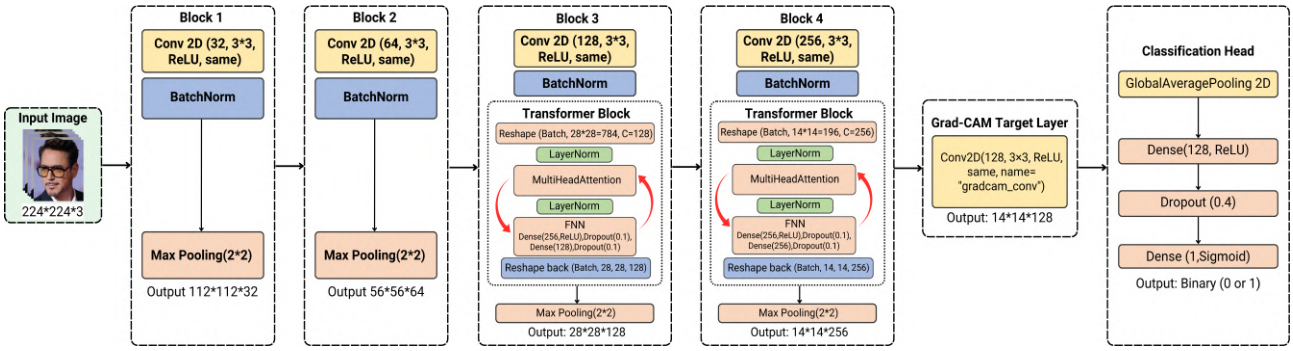

Fig. 2. The architecture of the proposed CNN–Transformer hybrid model, illustrates how input images are processed through convolutional blocks for spatial feature extraction, followed by Transformer layers to capture long-range dependencies. Grad-CAM is integrated to highlight discriminative regions, while the classification head performs binary prediction with interpretability.

dataset consisted of 98 well-known celebrities with a total of 12,216 authentic face images. After collecting the dataset, all individual directories were removed and merged into a single directory named "Real". Since the dataset initially contained only real images, deepfake samples were generated using a pre-trained `StyleGAN Ada2` model, producing the same number of synthetic images, 12,216, to ensure class balance. The final dataset, therefore, consisted of two categories, real and deepfake, with a total of 24,432 images. For experimentation, the dataset was split into 70% for training, 15% for validation, and 15% for testing, corresponding to 17,102 training images, 3,664 validation images, and 3,666 testing images, with equal distribution maintained across both classes. Fig. 3 shows sample images from the dataset, and Table I shows the data distribution across train, test, and validation.

(a) Real (b) Real (c) Real (d) Real (e) Real

(f) Fake (g) Fake (h) Fake (i) Fake (j) Fake

Fig. 3. Sample images representing Real and deepfake from the dataset.

TABLE I
DATA DISTRIBUTION ACROSS TRAIN, TEST, AND VALIDATION SETS.

Class	Training	Validation	Testing	Total
Real	8,551	1,832	1,833	12,216
DeepFake	8,551	1,832	1,833	12,216
Total	17,102	3,664	3,666	24,432

B. Data Preprocessing and Augmentation

The dataset contained multiple image formats such as JPG, PNG, and JPEG; therefore, all files were converted into the JPG format, which is more suitable for CNN and Transformer-based models. All images were uniformly resized to 224×224 pixels to ensure consistency in input dimensions for model training. To enhance model generalization, data augmentation was applied using `ImageDataGenerator`. The training generator included rescaling (1./255), brightness variation (0.8–1.2), random zoom (±0.1), horizontal flipping, and small width/height shifts (±0.1), simulating real-world variability. Validation and testing datasets were rescaled only, ensuring unbiased evaluation. The dataset was finally loaded into `TensorFlow.Keras` pipelines, preserving class balance and enabling efficient batch-wise training and assessment.

C. Proposed CNN-Transformer Model

We proposed a novel hybrid CNN-Transformer architecture for deepfake image classification, designed to capture both local texture details and long-range spatial dependencies characteristic of manipulated facial regions. Unlike transfer learning approaches, our network is trained entirely from scratch, with all parameters randomly initialized, allowing it to learn task-specific representations directly from the training data. The model accepts input images of size 224×224×3 and is composed of four sequential convolutional blocks, followed by two transformer encoder modules, a Grad-CAM target layer, and a compact classification head. The first, second, third, and fourth blocks use 32, 64, 128, and 256 filters, respectively, all with 3×3 kernels, ReLU activation, batch normalization, and max pooling for progressive spatial downsampling. Transformer encoder modules are integrated after the third and fourth blocks, enabling the model to capture global context and complex inter-pixel relationships at reduced spatial resolutions.

Each transformer encoder first applies layer normalization and multi-head self-attention with four heads, followed by a residual connection to preserve original features. The attention output is then passed through a feed-forward network with two dense layers (hidden dimension of 256) and dropout regularization, followed by another residual connection to stabilize learning and maintain gradient flow. This design allows the model to simultaneously exploit local convolutional features and global dependencies, which is crucial for detecting subtle, spatially distributed manipulations present in deepfake images.

After feature extraction, an additional convolutional layer with 128 filters, named `gradcam_conv`, is applied as the

target layer for Grad-CAM visualization, enabling the generation of interpretable class activation maps. The classification head then applies global average pooling, followed by a fully connected layer with 128 units and ReLU activation. A dropout rate of 0.4 is applied to prevent overfitting, and a final sigmoid neuron outputs the probability that the input image is real or fake. Therefore, we named the model TransCNN to reflect its hybrid design that integrates Transformer and CNN components for deepfake detection. Fig. 2 shows the architecture of the proposed CNN-Transformer Model.

IV. Experimental Setup and Implementation

The experiments were carried out on a system running Windows 11 with a 64-bit architecture, powered by an Intel® Core™ i5 8th Generation processor clocked between 1.60 GHz and 3.90 GHz, equipped with 12 GB of DDR4 RAM and an NVIDIA GeForce MX250 GPU with 2 GB of VRAM. For computationally intensive tasks, such as image preprocessing, model training, and evaluation, the Kaggle Notebook with dual NVIDIA Tesla T4 GPUs was utilized to accelerate computation. The end-to-end training and evaluation process for the CNN-Transformer model took over six hours to complete.

A. Training Setup

The proposed CNN-Transformer model was trained to classify images as either deepfake or real. All input images were resized to $224 \times 224 \times 3$ and normalized before training to ensure uniformity and improve convergence. The dataset was randomly split into training, validation, and test sets with a 70:15:15 ratio, ensuring class balance across all splits. To enhance generalization and reduce the risk of overfitting, data augmentation techniques were applied. The model was trained for 50 epochs using a batch size of 8, which was chosen to balance computational efficiency with stable convergence given the dataset size and available hardware resources.

B. Parameter Settings

The model was optimized using the Adam optimizer with an initial learning rate of 1×10^{-4}, and binary cross-entropy loss was employed to handle the binary classification task. Early stopping was applied with a patience of 15 epochs, automatically restoring the best-performing weights once validation performance ceased to improve. Additionally, a learning rate scheduler reduced the learning rate by a factor of 0.3 whenever the validation loss did not improve for 5 consecutive epochs. This parameter configuration ensured stable optimization, faster convergence, and improved generalization performance on unseen test data.

C. Evaluation Metrics

To assess the effectiveness of the proposed models, we utilized several well-established evaluation metrics, including the confusion matrix (CM), accuracy, precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve. The confusion matrix offers a clear representation of the model's predictive performance by summarizing the number of correctly and incorrectly classified samples for each class. The Table II presents the primary evaluation metrics along with their mathematical formulations. In binary classification scenarios, True Positives (TP) and True Negatives (TN) denote correctly predicted samples, whereas False Positives (FP) and False Negatives (FN) correspond to misclassified cases.

TABLE II
Performance Evaluation Metrics

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Represents the overall correctness of predictions; calculates the proportion of all correctly classified instances relative to total samples.
Precision	$\frac{TP}{TP+FP}$	Fraction of predicted positive cases that are actually true; reflects the reliability of positive predictions.
Recall	$\frac{TP}{TP+FN}$	Measures the model's capability to detect actual positive instances correctly.
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall; provides a balanced measure of performance considering both false positives and false negatives.

D. Model Interpretability with Grad-CAM

To improve interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize the regions of input images that most influenced the model's predictions. Grad-CAM works by computing the gradient of the predicted class score with respect to the final convolutional feature maps (`gradcam_conv` layer) and producing a heatmap that highlights discriminative regions. The heatmap was upsampled to the input image resolution and overlaid on the original image to generate visual explanations. This helps verify that the model focuses on relevant facial areas, such as eyes, mouth, and skin textures, which are often manipulated in deepfakes. These visualizations provide qualitative evidence of the model's reliability and support the trustworthiness of its classification results.

V. Experimental Result Analysis and Discussion

This section presents the experimental results of the proposed CNN-Transformer model for the classification of deepfake images. To assess its effectiveness, we compare its performance against a baseline CNN model that shares the same convolutional architecture but excludes the transformer encoder modules. The evaluation is conducted using widely adopted metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive assessment of both overall classification performance and the balance between false positives and false negatives. In addition to quantitative results, we also present model confidence scores alongside output images and Grad-CAM visualizations, which offer deeper insights into the model's decision-making process.

A. Result Analysis

Fig. 4 presents the training and validation performance curves of the proposed CNN-Transformer model over 50 epochs. The training accuracy consistently reaches 100% after the initial few epochs, while the validation accuracy fluctuates within the range of 97.5% to 98.5%, indicating strong generalization capability with minimal performance variation. The training loss steadily decreases and approaches zero, demonstrating effective model optimization. Meanwhile, the validation loss stabilizes around 0.05 after early fluctuations, reflecting that the model maintains a low and consistent error rate on unseen data. These results collectively confirm that the

proposed architecture achieves efficient convergence, avoids significant overfitting, and generalizes well to the validation set.

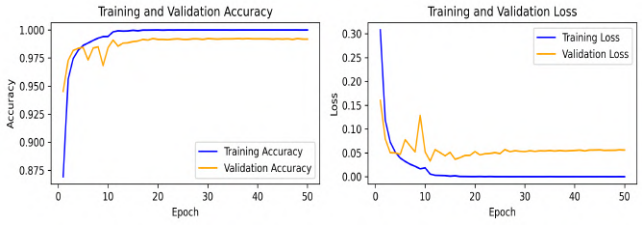

Fig. 4. Training and validation accuracy and loss curves of the proposed CNN-Transformer model.

Table III reports the performance of the baseline CNN model on the deepfake classification task. The model achieved a precision of 0.9868, a recall of 0.9814, an F1-score of 0.9841, and an accuracy of 0.9814 for the DeepFake class, indicating good capability to detect manipulated images. For the Real class, the model achieved a precision of 0.9816, a recall of 0.9869, an F1-score of 0.9842, and an accuracy of 0.9869, showing reasonably balanced performance between false positives and false negatives. The macro and weighted average metrics both achieved 0.9842, demonstrating stable performance across both classes.

TABLE III
PERFORMANCE OF BASELINE CNN FOR DEEPFAKE FORENSICS

Class	Precision	Recall	F1-score	Accuracy	Support
DeepFake	0.9868	0.9814	0.9841	0.9814	1833
Real	0.9816	0.9869	0.9842	0.9869	1833
Macro Avg	0.9842	0.9842	0.9842	0.9842	3666
Weighted Avg	0.9842	0.9842	0.9842	0.9842	3666

Table IV summarizes the classification performance of the proposed TransCNN model on the deepfake detection task. The model achieved consistently high scores across all metrics, with a precision of 0.9870, recall of 0.9913, F1-score of 0.9891, and accuracy of 0.9913 for the DeepFake class, indicating its strong ability to correctly identify manipulated images with minimal false positives. Similarly, for the Real class, the model attained a precision of 0.9912, a recall of 0.9869, an F1-score of 0.9891, and an accuracy of 0.9869, reflecting balanced performance in distinguishing authentic images. Both macro and weighted averages of precision, recall, and F1-score reached 0.9891, demonstrating that the model performs consistently across both classes and is not biased toward either. These results collectively highlight the effectiveness of integrating transformer modules into the CNN backbone, enabling the model to capture both local texture patterns and global contextual information, which leads to superior performance in deepfake classification.

TABLE IV
PERFORMANCE OF CNN-TRANSFORMER FOR DEEPFAKE FORENSICS

Class	Precision	Recall	F1-score	Accuracy	Support
DeepFake	0.9870	0.9913	0.9891	0.9913	1833
Real	0.9912	0.9869	0.9891	0.9869	1833
Macro Avg	0.9891	0.9891	0.9891	0.9891	3666
Weighted Avg	0.9891	0.9891	0.9891	0.9891	3666

Fig. 5 illustrates the confusion matrix for the proposed TransCNN model. The model correctly classified 1817 deepfake samples and 1809 real samples, achieving very low misclassification rates with only 16 deepfakes misclassified as real and 24 real samples misclassified as deepfakes. These results correspond to a high overall accuracy, balanced precision, and recall, demonstrating the model's robustness and ability to generalize well across both classes.

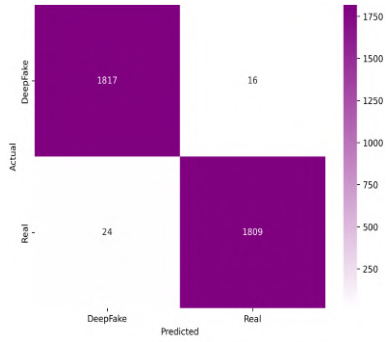

Fig. 5. Confusion matrix of the proposed CNN-Transformer model

Fig. 6 presents the Receiver Operating Characteristic (ROC) curve of the proposed TransCNN model. The model achieves an Area Under the Curve (AUC) of 1.00 for both real and fake classes, indicating near-perfect discrimination capability. The ROC curve closely aligns with the top-left corner, demonstrating an almost zero false positive rate and a true positive rate approaching one.

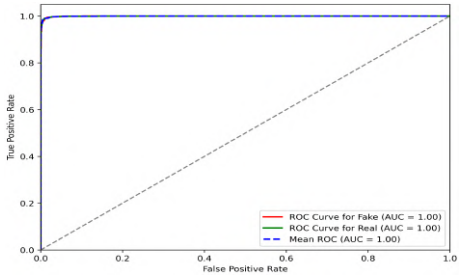

Fig. 6. ROC curve of the proposed CNN-Transformer model

Figure 7 illustrates the qualitative results of the proposed TransCNN model for deepfake detection, showing both correctly and confidently classified images. The model consistently identifies real and deepfake faces with perfect confidence scores, 1.00 for real faces, 0.00 for deepfakes, across diverse facial appearances, lighting conditions, and background settings. Real images are classified as "Real" with maximum confidence, while manipulated (deepfake) images are labeled as "DeepFake" with zero confidence toward the real class.

Figure 8 illustrates the Grad-CAM visualizations generated by the proposed TransCNN model for deepfake detection. Each row presents the original image, its corresponding Grad-CAM heatmap, and the overlay visualization. The heatmaps highlight the discriminative regions that contributed most to the model's decision, revealing that the network effectively focuses on facial regions such as eyes, mouth, and surrounding textures, where manipulations are typically

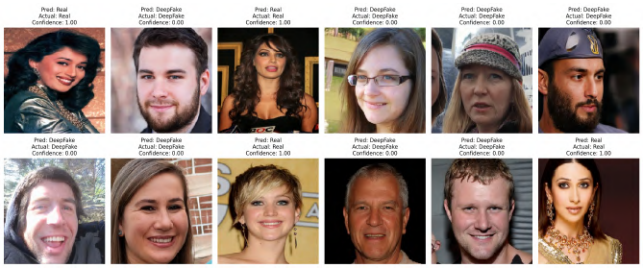

Fig. 7. Sample classification results of the proposed CNN-Transformer model with corresponding confidence scores, showcasing perfect detection of both real and deepfake images across diverse samples.

present. The overlays demonstrate that the model attends to subtle artifacts and inconsistencies in the manipulated images, providing interpretability and transparency into the decision-making process. These visualizations confirm that the proposed architecture is capable of capturing both local texture anomalies and global spatial relationships, leading to improved classification performance.

Fig. 8. Grad-CAM visualizations of the proposed CNN-Transformer model, highlighting key facial regions used to distinguish real and deepfake images.

B. Discussion and Limitations

The experimental results demonstrate that the proposed TransCNN architecture consistently outperforms the baseline CNN model across all evaluation metrics. Although the improvement over the baseline CNN appears modest, it holds significant value given the highly sensitive nature of deepfake detection. Even minor performance gains can have substantial implications in real-world applications, where a small number of misclassifications could lead to serious ethical, social, and legal consequences. Furthermore, we intentionally chose not to compare our approach with state-of-the-art deepfake detection methods, as differences in datasets, preprocessing pipelines, and evaluation protocols would make such comparisons less meaningful and potentially misleading. However, this work has several limitations. First, the model was trained and evaluated on a single dataset, which may limit its generalizability to unseen, real-world deepfakes with diverse generation techniques and compression levels. Second, the Transformer block introduces additional computational overhead, which could hinder deployment in resource-constrained or real-time environments. Third, the study does not analyze the robustness of the model against adversarial attacks or perturbations, which is a growing concern in deepfake detection.

VI. Conclusion and Future Work

This research presented a hybrid CNN-Transformer architecture for deepfake image classification, trained from scratch to learn dataset-specific features without relying on pretrained weights. The integration of Transformer encoder blocks after the deeper convolutional layers enabled the model to capture both local texture information and global spatial dependencies, resulting in improved performance over the baseline CNN across all evaluation metrics. Grad-CAM visualizations confirmed that the model attends to semantically meaningful facial regions, enhancing interpretability and trust in its predictions. Future work will focus on extending the evaluation to cross-dataset settings to assess generalizability to real-world deepfakes, optimizing the Transformer module for deployment in resource-constrained environments, and investigating robustness against adversarial manipulations to ensure reliability in practical applications.

References

[1] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: an overview," in *Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020*, pp. 557–566, Springer, 2021.

[2] D. Chapagain, N. Kshetri, and B. Aryal, "Deepfake disasters: A comprehensive review of technology, ethical concerns, countermeasures, and societal implications," in *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, pp. 1–9, IEEE, 2024.

[3] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, and V. Vimal, "Deepfake generation and detection: Case study and challenges," *IEEE Access*, vol. 11, pp. 143296–143323, 2023.

[4] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov, and A. S. Smirnov, "Methods of deepfake detection based on machine learning," in *2020 IEEE conference of russian young researchers in electrical and electronic engineering (EIConRus)*, pp. 408–411, IEEE, 2020.

[5] M. S. Hossen, M. Saiduzzaman, P. Shaha, and M. K. Nasir, "Jellyfish species identification: A cnn-based artificial neural network approach," in *Proceedings of the International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, IEEE, 2025.

[6] N. Celebi and Q. Liu, "Spatiotemporal deepfake video detection: A hybrid cnn-transformer approach with frequency analysis," in *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*, pp. 216–221, IEEE, 2025.

[7] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, "Deepfake image forensics for privacy protection and authenticity using deep learning," *Information*, vol. 16, no. 4, 2025.

[8] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, "A hybrid deep learning framework for deepfake detection using temporal and spatial features," *IEEE Access*, vol. 13, pp. 79560–79570, 2025.

[9] I. Ambreen, M. Aatif, Z. Jalil, F. Iqbal, and A. Marrington, "Pvit: A hybrid model for deepfake face detection using patch vision transformers and deep learning," in *2025 12th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pp. 58–66, 2025.

[10] S. A. Kumar, M. A. Kumar, R. R. Vincent, U. U. Hegde, A. M. S, and N. N. Pasha, "Deepfake image detection using deep learning," in *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 492–497, 2025.

[11] D. Awasthi, P. Khare, V. K. Srivastava, A. K. Singh, and B. B. Gupta, "Deepnet: Protection of deepfake images with aid of deep learning networks," *Image and Vision Computing*, vol. 158, p. 105540, 2025.

[12] A. Jaiswal, A. Chamatkar, A. P. Kharat, B. Sharma, I. B. Dhaou, and S. A. Chaurasia, "A novel framework for deepfake image detection using deep learning approach," in *2025 22nd International Learning and Technology Conference (L&T)*, vol. 22, pp. 269–273, 2025.

[13] R. Sharma, T. Sharma, S. Arora, and S. Verma, "A comparative analysis of deep neural network models for deepfake image detection generated using ai," in *2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, pp. 1–5, 2025.

[14] K. Magoo, T. C. Amity, R. Gandhi, and M. Sharma, "Deepfake image and forged signature detection using machine learning," in *2025 International Conference on Engineering, Technology & Management (ICETM)*, pp. 1–6, 2025.

[15] H. Soni, "Celebrity images for face recognition." Kaggle, Available at: https://www.kaggle.com/datasets/hemantsoni042/celebrity-images-for-face-recognition, 2023.