IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 May 2025, accepted 4 June 2025, date of publication 10 June 2025, date of current version 18 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3578343

RESEARCH ARTICLE

Attention-Enhanced CNN for High-Performance Deepfake Detection: A Multi-Dataset Study

SUBHRAM DASGUPTA, KUSHAL BADAL, SWETHA CHITTAM, MD TASNIM ALAM, AND KAUSHIK ROY, (Senior Member, IEEE)
Department of Computer Science, North Carolina A&T State University, Greensboro, NC 27411, USA

Corresponding author: Subhram Dasgupta (sdasgupta@aggies.ncat.edu)

This work was supported by the National Science Foundation (NSF) under Award 1900187.

ABSTRACT Deepfakes, which emerge from advanced deep learning techniques, present complex ethical and security challenges across media and communication landscapes. While offering creative potential in education and entertainment, synthetic media technologies simultaneously threaten societal trust through potential misinformation, opinion manipulation, privacy violations, and identity fraud. With the advancement of deep learning models, creating deepfake images has become easier and more convincing, resulting in the development of reliable deepfake detection models. This research works with a method that combines multi-head self-attention (MHSA) with a custom-designed convolutional neural network (CNN) to develop a robust deepfake detection model. We created a dataset called the Center for Cyber Defense DeepFake (CCDDF) dataset by generating fake images using publicly available Artificial Intelligence (AI) tools and trained our model on these data, achieving a detection accuracy of 97% and an AUC score of 99.58. Additionally, we evaluated our model on the 140K Real and Fake Faces dataset and the Celeb-DF v2 dataset, where it demonstrated exceptional performance with accuracies of 98% and 94% respectively, and corresponding AUC scores of 99.75 and 98.72. We utilized attention heatmap visualizations to analyze the model's decision-making process to enhance qualitative interpretability. Our results demonstrate the effectiveness of combining multi-head self-attention with a convolutional neural network for deepfake detection, highlighting its strong performance across multiple datasets and its potential for real-world applications.

INDEX TERMS Deepfake detection, convolutional neural networks (CNNs), multi-head self attention (MHSA), attention heatmap visualization, hybrid CNN-attention model.

I. INTRODUCTION

The rapid evolution of AI has given rise to an intriguing yet potentially concerning phenomenon known as 'deepfakes.' These synthetic media creations, whose name derives from the fusion of 'deep learning' and 'fake,' represent a sophisticated application of AI technology that can convincingly replicate a person's appearance, voice, or actions. As computational power and AI capabilities have advanced, deepfakes have transformed from experimental curiosity to a widespread technological phenomenon, bringing both promising opportunities and significant societal challenges [1]. Deepfakes are generally made with Generative Adversarial Networks (GANs), which are composed of a generator and discriminator [2]. The generator produces synthetic content by learning patterns from data such as facial expressions. In contrast, the discriminator assesses the authenticity of the generated material by comparing it to authentic images and provides input to help the generator improve its output. These networks improve the quality of the generated media through iterative training until they are nearly identical to the actual content [3]. Other methods include the use of autoencoders, in which an encoder compresses the input data and a decoder reconstructs them to simulate real-world scenarios [4]. The applications of deepfake technology span a broad spectrum, from beneficial innovations to potentially harmful misuses. On the positive side, researchers have developed interactive digital twins that can serve as human substitutes in cyber-physical systems [5]. The entertainment industry has embraced deepfakes to create sophisticated visual effects, whereas language learning platforms use them to synchronize

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

lip movements for seamless content dubbing [6], [7], [8]. However, this technology has also been used to disseminate misinformation, often manifesting as viral social media content featuring manipulated videos of public figures [9]. More concerning are malicious applications, including the creation of non-consensual intimate content and attempts at political manipulation using fabricated videos [1].

To address these challenges, researchers have developed various deepfake detection methods that are, broadly categorized into detection and prevention approaches. Our research focuses specifically on spatial artifact-based detection, emphasizing on human image manipulation. Several key inconsistencies in spatial artifacts provide crucial indicators for deepfake detection [1].

Our research experiments combine a sequential convolutional neural network (CNN) architecture with Multi-Head Self-Attention (MHSA) for deepfake detection [10]. While CNN excels in extracting local features, the self-attention mechanism enhances the model's capability to identify crucial features and capture long-range dependencies that might elude traditional CNN analysis [10]. Although existing approaches, such as Vision Transformers (ViTs) employ multi-head attention, they often require substantial computational resources and training overhead [11]. Our methodology aims to achieve superior detection accuracy while maintaining computational efficiency. The integration of MHSA mechanisms (proposed in [10]) into CNN architectures represents a significant advancement in feature extraction capabilities. This combination allows our model to capture global contextual information dynamically while emphasizing relevant features, resulting in more effective and efficient data processing [10]. The MHSA component acts as an intelligent filter, amplifying informative features while suppressing less significant ones, leading to a more robust feature extraction from the input data. Our experimental results demonstrate that this streamlined architecture, combining CNN with MHSA, achieves comparable performance to state-of-the-art methodologies while potentially offering better computational efficiency.

A. RESEARCH CONTRIBUTION

The following are the significant contributions of this study:

- Experimental CNN-MHSA hybrid architecture for deepfake detection that combines local feature extraction with attention mechanisms
- Comprehensive model evaluation using both custom-generated synthetic images and two benchmark deepfake datasets
- Model interpretability analysis through heatmap visualizations of prediction outputs

B. ARTICLE LAYOUT

The remainder of this paper is organized as follows. Section II discusses the existing approaches for deepfake detection. Section III presents the data collection and preprocessing steps and the methodologies used for the problem. Section IV discusses the experimental results and performance evaluation of the model based on various metrics, and Section V presents the conclusion of our work.

II. RELATED WORKS

The following section discusses existing studies in the domain of deepfake detection. This section is subdivided into preliminary work II-A and existing challenges II-B, followed by a comprehensive table summarizing the deepfake detection works in Table 1.

A. PRELIMINARY WORK

Traditional machine learning (ML) approaches offer distinct advantages in deepfake detection owning to their inherent interpretability and straightforward parameter-tuning capabilities. Tree-based algorithms, such as Decision Trees, Random Forests, and Extremely Randomized Trees, provide transparent decision processes, making them particularly valuable for understanding detection mechanisms. These methods have proven to be effective in identifying artificial patterns generated by GANs in synthetic media. Early ML-based detection methods focused on identifying specific facial manipulations, such as alterations in eye shading or ear features. However, these single-feature approaches have proven to be limited in their detection capabilities. To address this limitation, researchers have developed comprehensive methods. Research advances in ML-based deepfake detection have evolved from single-feature analysis to more sophisticated approaches. [12] pioneered multi-feature detection by combining various facial characteristics, whereas [13], [14], [15] enhanced detection through biological signal analysis across spatial-temporal dimensions using facial landmarks. Further innovations include 3D head pose estimation [16] and Habeeba et al.'s efficient MLP model which analyzes facial artifacts with minimal computation [17]. One significant approach examines blending artifacts that occur when the generated content is integrated back into the frame, leading to the development of CNN models that utilize edge detection techniques [18]. Another promising strategy analyzes the relationship between facial features and their surrounding context, as demonstrated by Nirkin et al.'s framework, which not only contrasts foreground and background information but also enables networks to automatically identify discriminative features [19]. Additionally, GAN-generated deepfakes often leave distinctive fingerprints that can be identified through a careful analysis of subtle patterns and features [1]. Although these ML methods can achieve up to 98% accuracy, their performance varies significantly based on the dataset characteristics and feature selection. Testing on heterogeneous datasets often reduces accuracy to approximately 50%, underscoring the challenge of developing generalizable detection systems.

Deep-learning approaches have dominated recent advances in deepfake detection research. Initial methods focused on identifying GAN-generated artifacts [20] and extracting features from RGB data [21], [22], while subsequent work incorporated physiological measurements such as heartbeat patterns [23]. A significant milestone was the development of inception module-based networks (Meso-4 and MesoInception-4) [24], which established the effectiveness of deep learning in video-based detection. Research has shown that deep CNNs consistently outperform shallow architectures [25], [26], [27] in supervised scenarios. Feature extraction has emerged as a crucial focus, with researchers developing methods for analyzing handcrafted features [28], spatiotemporal patterns [29], and facial landmarks [30]. Innovation continued with the introduction of capsule networks [31], [32] and ensemble learning techniques achieving accuracy rates exceeding 99% [33]. RNN-based approaches [34] have advanced the field by enabling feature extraction at both the micro and macroscopic levels. To address the overfitting challenges, researchers have developed optical flow techniques [35] and autoencoder architectures [36], [37]. Recent developments have significantly expanded detection capabilities through multimodal analysis combining audio and visual features [38]. The introduction of attention-based architectures has further revolutionized this domain, enabling more nuanced and context-aware detection mechanisms [39], [40]. Although these approaches have demonstrated impressive results, the field continues to grapple with challenges in model generalization and robustness across diverse datasets.

B. CHALLENGES

Prior research on deepfake detection has encountered several fundamental challenges that continue to shape the trajectory of current research. Recent studies have highlighted the persistent difficulty in achieving robust generalization across diverse deepfake variants. Although contemporary approaches incorporate sophisticated neural architectures, including Convolutional Neural Networks (CNNs) and attention mechanisms, the rapid evolution of deepfake generation techniques frequently outpaces detection capabilities. For instance, Wang et al. [41] demonstrated that traditional CNN-based models, despite employing extensive data augmentation strategies, struggle to identify novel manipulation patterns that deviate from their training distributions. The challenges of dataset quality and representation have emerged as critical bottlenecks in advancing deepfake detection research. Rossler et al. [25] revealed that publicly available datasets often exhibit limited diversity in manipulation techniques, potentially leading to models that overfit specific artifact patterns rather than learning generalizable detection features. This limitation is particularly pronounced in binary classification approaches, which, although computationally efficient, may oversimplify the complex spectrum of deepfake manipulations present in real-world scenarios. Feature extraction and model interpretability are significant research challenges. Although attention-based architectures have shown promise in improving feature localization, as demonstrated by Zhao et al. [10], the field still grapples with effectively identifying and explaining fine-grained manipulation artifacts. This challenge becomes particularly acute when models are trained on datasets with limited samples or when dealing with sophisticated deepfake generation techniques that produce nearly imperceptible alterations. Growing concern regarding adversarial attacks has introduced additional complexity to the detection landscape. Studies by Li et al. [42] demonstrated that current detection models remain vulnerable to carefully crafted adversarial perturbations, post-processing modifications, and compression artifacts. Although data augmentation and regularization techniques offer protection, developing robust defenses against adversarial manipulation continues to be an active area of research. Computational efficiency in real-world applications presents a practical challenge that intersects with theoretical advances. Although state-of-the-art models achieve impressive accuracy rates, their deployment in real-time systems often requires significant computational resources. Nguyen et al. [6] highlighted the inherent trade-off between model complexity and inference speed, particularly when implementing attention mechanisms along with traditional CNN architectures. Table 1 provides a summary of the existing work on deepfake detection.

III. METHODOLOGY

Deepfake content poses an increasing threat to digital media authenticity; however, current state-of-the-art detection methods often require substantial computational resources owning to their complex parallel architectures and dense interconnections, limiting their practical deployment. This study aims to develop a memory-efficient deepfake detection framework through the sequential integration of CNN and MHSA mechanisms (MHSA proposed in [55] is combined with Deep Learning (DL) models [10] to leverage both local feature extraction capabilities and global contextual understanding for enhanced detection accuracy. In this research, we modified the traditional attention mechanism by implementing a lightweight sequential architecture), optimizing resource utilization while maintaining robust detection performance through effective feature extraction and spatial relationship modeling. This section presents the data collection and preprocessing steps and the architectural framework of the proposed approach. This study proposes a CNN-MHSA integration for robust deepfake detection. Our model employs 696,001 parameters (2.66 MB) is shown in Table 2, representing a streamlined architecture compared to many contemporary approaches. The sequential integration design, where the multi-head attention mechanism follows convolutional feature extraction, allows the model to benefit from attention-based contextual modeling while maintaining a focused architectural design.

Our primary objective in this study was to develop and evaluate a detection framework that performs effectively across diverse deepfake datasets, which our experimental results demonstrate through high accuracy and AUC scores on the Celeb-DF v2, Custom Center for Cyber Defense DeepFake (CCDDF), and 140K datasets.

TABLE 1. Summary of deepfake detection techniques.

Model	Dataset	Methods	Metrics
CNN + SVM [43]	140k [44]	Convolution + SVM	Accuracy, AUC
Light-weight CNN [45]	140k [44]	Efficient CNN Architecture	Accuracy
MLP + LSTM [46]	140k [44]	Multilayer Perceptron + LSTM	Accuracy
LightFFDNets v2 [47]	140k [44]	Lightweight Deep Network	Accuracy
Custom CNN model [48]	140k [44]	Custom Convolutional Model	Accuracy
CNN + SE [40]	DFFD [49]	CNN enhanced with SE block	Accuracy, AUC
Defakehop [50]	Celeb-DF v2 [42]	Successive Subspace Learning	AUC
Fusion Network [51]	Celeb-DF v2 [42]	Multi-stream Fusion Network	AUC

A. DATA COLLECTION

The experimental validation of our framework was conducted across three distinct datasets, each chosen to evaluate specific aspects of the model's performance. Our initial experiment utilized the CCDDF dataset we created, comprising synthetic and authentic facial images. We generated 5,000 synthetic images via the generated.photos platform using automated scripts for this dataset while sourcing an equivalent number of real photographs from the FFHQ dataset [52]. The CCDDF dataset used in this study is publicly available through IEEE DataPort [53].To enhance the training process, we performed data augmentation on both classes, expanding each to 8,000 images, resulting in a balanced dataset of 16,000 samples.

The second phase of the experiment employed a large-scale dataset sourced from Kaggle, encompassing 140,000 images equally divided between authentic and synthetic faces [44]. The synthetic portion was created using StyleGAN technology, whereas authentic images were sourced from the FFHQ dataset. This extensive collection enables a thorough assessment of the capability of our model to handle diverse facial characteristics and generation patterns at scale.

Our final experimental phase utilized the Celeb-DF v2 dataset [42], which presents a unique challenge because of its class imbalance. This collection comprises manipulated celebrity videos alongside their original counterparts sourced from YouTube. The dataset's composition of real celebrity footage and its corresponding deepfake versions provided an ideal testbed for evaluating our model's performance against high-quality manipulated content under varying conditions.

We conducted separate experiments on each dataset to ensure thorough evaluation across different data distributions and manipulation techniques. This method enabled us to validate the consistency and effectiveness of our model across diverse scenarios. The distribution of images across classes for each dataset is illustrated in Figure 1, while the representative samples from our experiments are shown in Figure 2.

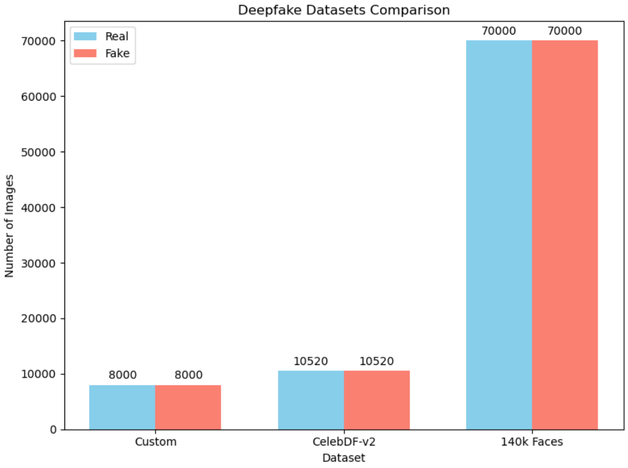


FIGURE 1. Total Number of images in each class for the datasets used in the experiments.

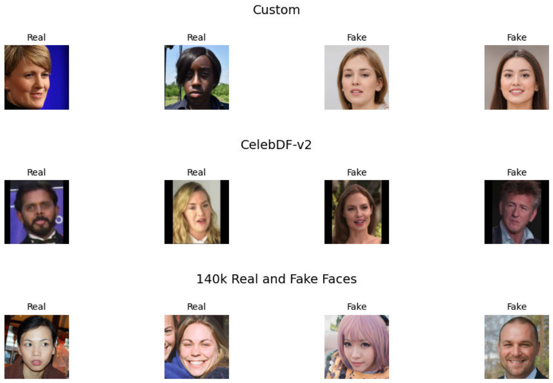


FIGURE 2. Samples of the data used in the experiments [42], [44], [53].

B. DATA PREPROCESSING

Our data preprocessing methodology focuses on standardizing diverse datasets while preserving their unique characteristics for effective model evaluation. The Celeb-DF v2 dataset presents unique challenges because it contains video content rather than static images. We extracted facial regions from these videos using Multi-task Cascaded Convolutional Networks (MTCNN) [54], implementing a strategic sampling approach to address the inherent class imbalance between real and fake videos and extracting more frames from real videos to create a more balanced training set.

The Celeb-DF v2 dataset exhibited a lower resolution than other datasets in our study. To compensate for this limitation, we modified our face extraction parameters to include additional contextual information around the facial region. This modification proved beneficial as the surrounding visual information often contained subtle artifacts that could aid in deepfake detection. After face extraction, we normalized all images across the three datasets to a uniform dimension of 224 × 224 pixels to maintain consistency in model input size.

We further refined our dataset using targeted data augmentation techniques to address any remaining class imbalances. These augmentations were specifically chosen to preserve the critical facial features while introducing meaningful variations in the training data. Through this comprehensive preprocessing approach, we successfully standardized our diverse datasets into a uniform format while retaining their distinctive characteristics, thereby enabling a thorough evaluation of our detection framework across various conditions.

C. CNN ARCHITECTURE

Our model employs a custom-designed sequential CNN architecture rather than a pre-trained model or established architecture (such as VGG, ResNet, or Inception). This custom approach was chosen to optimize the balance between computational efficiency and feature extraction capabilities, which was critical for the subsequent integration with our Multi-Head Self-Attention (MHSA) mechanism. The custom architecture allows us to precisely control network depth and feature dimensionality at each layer, ensuring compatibility with the MHSA component while maintaining memory efficiency.

The CNN component of our architecture serves as the primary feature extractor and is designed to capture both the local and hierarchical patterns in the input data. The network architecture begins with an input layer that accepts image data of dimensions 224 × 224 × 3. The network's core consists of four sequential convolutional blocks with progressively increasing filter sizes of 32, 64, 128, and 256 filters, respectively. This gradual expansion in the number of filters enables the network to learn increasingly complex feature representations at different scales. Each convolutional block in our architecture implements a systematic combination of layers to ensure optimal feature extraction and model regularization. The convolutional layers utilize 3 × 3 kernels initialized using the He normal initialization scheme, which helps maintain stable gradient flow during training. To prevent overfitting, we incorporate L2 kernel regularization with a coefficient of 1e-4. Following each convolution operation, a batch normalization layer is applied to stabilize the learning process and accelerate training. The normalized features are then passed through a ReLU activation function to introduce non-linearity. Spatial dimensionality reduction is achieved through max-pooling layers with a pool size of 2 × 2, effectively halving the spatial dimensions while retaining the most salient features. A dropout layer is included after each block to further enhance the model's generalization capabilities. The features extracted by the convolutional blocks are flattened into a one-dimensional vector before being fed into a series of dense layers. The dense layer configuration consists of two fully connected layers with 128 and 64 neurons, respectively, providing progressive dimensionality reduction while maintaining essential feature relationships. The network culminates in a final output layer designed for binary classification, distinguishing between real and fake images. A diagrammatic representation of the CNN module of the full architecture is given in Figure 3.

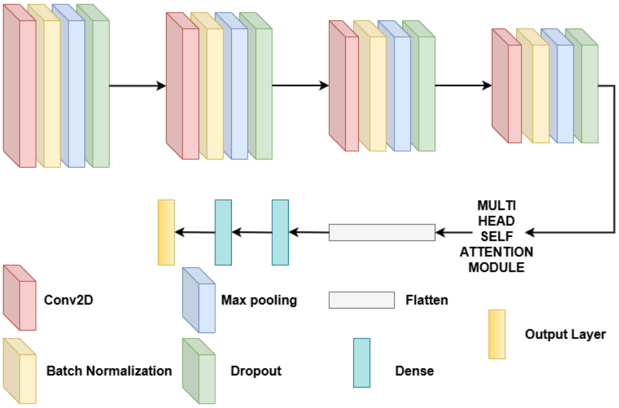


FIGURE 3. Diagrammatic representation of the CNN module.

D. MHSA MECHANISM

Following the fourth convolutional block, we implemented an MHSA mechanism to capture long-range dependencies and contextual relationships within the extracted feature space. The feature maps from the final convolutional layer, which contained 256 channels, were reshaped to prepare them for the attention mechanism. This reshaping operation flattens the spatial dimensions while preserving channel information, resulting in a sequence of feature vectors that can be processed by the attention mechanism. The MHSA module employed four attention heads, each operating with a key dimension of 64. This configuration allows the model to capture different types of relationships and patterns in parallel, because each head can potentially focus on different aspects of the feature space. To prevent overfitting and ensure robust learning, we incorporated dropout mechanisms at multiple stages: a dropout rate of 0.3 is applied before the attention operation, and an additional dropout rate of 0.2 is implemented within the attention mechanism itself. To maintain the network's ability to preserve important low-level features, we implemented a residual connection that adds the attention output back to the input features. This skip connection helps mitigate the potential loss of information during the attention operation and facilitates a better gradient flow during training. The combined features are then normalized using layer normalization, which helps stabilize the training process by normalizing the activations across the feature dimensions. Following the attention mechanism, the features are aggregated using global average pooling, which reduces the spatial dimensions while maintaining learned feature relationships. This pooled representation serves as a compact, informative feature vector that captures both the local patterns from the CNN and long-range dependencies from the attention mechanism.

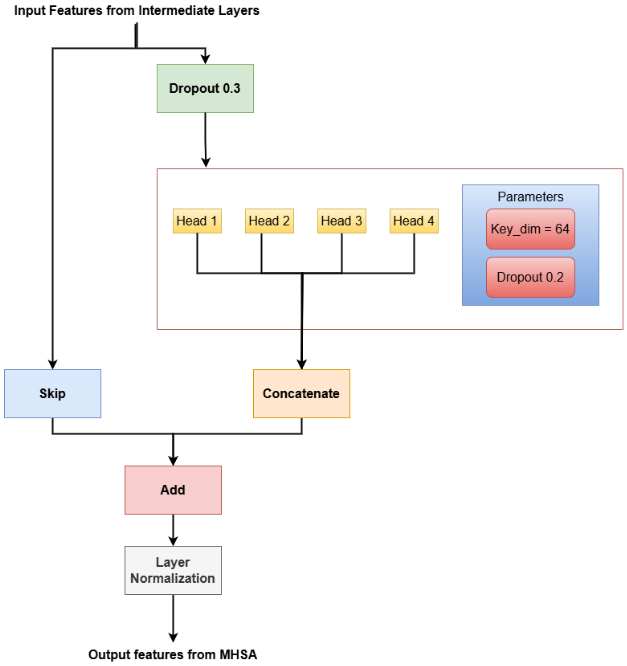


FIGURE 4. Diagrammatic representation of the MHSA module.

Figure 4. shows a diagrammatic representation of the MHSA module.

The mathematical formulation of the MHSA module, as described in [55] and used in our model, is given as follows:

$$\mathrm{MHA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O, \quad (1)$$

where each attention head is computed as:

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V). \quad (2)$$

The scaled dot-product attention is given by:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where $Q \in \mathbb{R}^{n\times d}$, $K \in \mathbb{R}^{n\times d}$, and $V \in \mathbb{R}^{n\times d}$ are the query, key, and value matrices, respectively; and d_k is the key dimension.

A residual connection and layer normalization are applied:

$$X' = \mathrm{LayerNorm}(X + \mathrm{MHA}(X, X, X)). \quad (4)$$

E. ATTENTION HEATMAP VISUALIZATION

To provide interpretability and insights into the model's decision-making process, we implemented an attention heatmap visualization methodology that reveals the regions of focus during image classification. The visualization process extracts intermediate outputs from the final convolutional layer and the multi-head attention layer through an intermediate model, enabling the analysis of both spatial features and attention weights. For each input image, we compute a spatial attention map by averaging the attention weights across all attention heads, followed by normalization to the range [0, 1]. The normalized attention map is then resized to match the original input image dimensions using bilinear interpolation and overlaid on the original image using a color-coded scheme with controlled transparency. This visualization approach is applied to both correctly and incorrectly classified images, enabling direct observation of the model's attention patterns and providing insights into the regions that influence classification decisions. The resulting heatmaps offer valuable insights into the model's behavior, highlighting areas that contribute to successful classification or potential misclassification, thereby enhancing our understanding of the model's decision-making process.

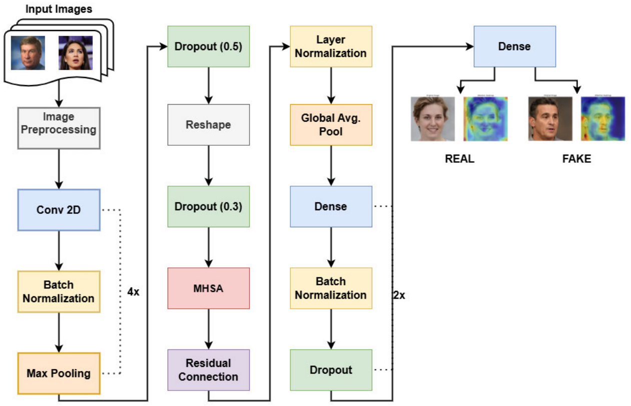


FIGURE 5. Proposed architecture.

F. DESCRIPTION OF THE PROPOSED MODEL ARCHITECTURE

We present a novel deepfake detection architecture that strategically combines Convolutional Neural Networks (CNN) with Multi-Head Self-Attention (MHSA) mechanisms. Our design introduces MHSA after the fourth CNN layer, enabling sophisticated contextual feature extraction while maintaining low computational overhead. We evaluated this architecture across three diverse datasets: a custom dataset built from FFHQ images, the 140_K Real and Fake Faces dataset, and CelebDF-V2 dataset. Through extensive testing, our model achieved classification accuracy and AUC scores comparable to those of state-of-the-art approaches while significantly reducing memory requirements. The complete model architecture is illustrated in Figure 5.

The model processes RGB input images through a series of four convolutional layers, each employing increasing filter sizes to capture the hierarchical features. These features were refined using batch normalization and max pooling operations before being processed by a 4-headed self-attention mechanism. This unique integration of MHSA enables our model to simultaneously analyze local spatial patterns and global image dependencies. The final classification was achieved through a series of dense layers that output a binary

TABLE 2. Model architecture summary.

Layer (Type)	Output Shape	Param #	Connected to
Input Layer (InputLayer)	(None, 224, 224, 3)	0	-
Conv2D	(None, 224, 224, 32)	896	input_layer[0][0]
BatchNormalization	(None, 224, 224, 32)	128	conv2d[0][0]
MaxPooling2D	(None, 112, 112, 32)	0	batch_normalization[0][0]
Conv2D	(None, 112, 112, 64)	18,496	max_pooling2d[0][0]
BatchNormalization	(None, 112, 112, 64)	256	conv2d_1[0][0]
MaxPooling2D	(None, 56, 56, 64)	0	batch_normalization_1[0][0]
Conv2D	(None, 56, 56, 128)	73,856	max_pooling2d_1[0][0]
BatchNormalization	(None, 56, 56, 128)	512	conv2d_2[0][0]
MaxPooling2D	(None, 28, 28, 128)	0	batch_normalization_2[0][0]
Conv2D	(None, 28, 28, 256)	295,168	max_pooling2d_2[0][0]
BatchNormalization	(None, 28, 28, 256)	1,024	conv2d_3[0][0]
MaxPooling2D	(None, 14, 14, 256)	0	batch_normalization_3[0][0]
Dropout	(None, 14, 14, 256)	0	max_pooling2d_3[0][0]
Reshape	(None, 196, 256)	0	dropout[0][0]
Dropout	(None, 196, 256)	0	reshape[0][0]
MultiHeadAttention	(None, 196, 256)	263,168	dropout_1[0][0], dropout_1[0][0]
Add (Residual Connection)	(None, 196, 256)	0	dropout_1[0][0], multi_head_attention[0][0]
LayerNormalization	(None, 196, 256)	512	add[0][0]
GlobalAveragePooling1D	(None, 256)	0	layer_normalization[0][0]
Dense	(None, 128)	32,896	global_average_pooling1d[0][0]
BatchNormalization	(None, 128)	512	dense[0][0]
Dropout	(None, 128)	0	batch_normalization_4[0][0]
Dense	(None, 64)	8,256	dropout_3[0][0]
BatchNormalization	(None, 64)	256	dense_1[0][0]
Dropout	(None, 64)	0	batch_normalization_5[0][0]
Dense (Output Layer)	(None, 1)	65	dropout_4[0][0]
Total Parameters	696,001 (2.66 MB)		
Trainable Parameters	694,657 (2.65 MB)		
Non-trainable Parameters	1,344 (5.25 KB)		

decision that distinguishes between authentic and synthetic images. A detailed summary of the model architecture is provided in Table 2.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we evaluate the performance of the proposed CNN-MHSA architecture using three distinct experimental datasets. Our analysis focused on both quantitative and qualitative insights of the model's detection capabilities. The custom dataset experiment validated our model's ability to identify AI-generated images from generated.photos AI tool against authentic FFHQ images, thereby establishing a baseline for synthetic image detection. The evaluation on the larger 140_K Real and Fake Faces dataset demonstrated the scalability and robustness of our model in handling diverse facial features and StyleGAN-generated images. The challenging CelebDF-V2 dataset provided insights into the effectiveness of our model in detecting video-based deepfakes despite variations in image quality and lighting conditions. Notably, our sequential integration of MHSA after the fourth CNN layer achieved competitive performance while significantly reducing memory overhead compared to conventional parallel architectures. The attention visualization maps, shown in Figure 12, Figure 13, and Figure 14 reveal that our model effectively identifies discriminative facial regions and potential manipulation artifacts, particularly focusing on areas around the eyes, mouth, and facial boundaries where synthetic patterns are most prevalent. These results demonstrate that our architecture successfully balances the computational efficiency with the detection accuracy across various datasets and generation techniques.

A. EXPERIMENTAL SETUP

Our experimental framework was implemented on a high-performance computing system running the Ubuntu 22.04 LTS (64-bit) operating system. The hardware configuration consisted of dual NVIDIA GeForce RTX 4090 graphics cards, each equipped with 24 GB of VRAM, enabling the efficient parallel processing of deep learning operations. The system was powered by an AMD Ryzen Threadripper Pro processor with 134.9 GB RAM, providing ample memory resources for handling large-scale datasets and complex model architectures. This robust computational environment ensured efficient training and evaluation of our proposed CNN-MHSA framework across multiple datasets while maintaining stable performance during extensive experimentation.

Our model uses a specialized version of focal loss, which improves upon regular binary cross-entropy by focusing more on hard-to-classify examples. We set it up with two main control knobs: an alpha value of 0.25 to help with unbalanced classes, and a gamma value of 2 to control how much we focus on the tricky cases. The function operates in an intelligent manner and first assesses the model's confidence in each prediction by using binary cross-entropy. Then, it adjusts the loss value such that when the model is very confident (but wrong), it becomes penalized more heavily. This helps prevent the model from becoming complacent with 'easy' examples, and keeps it focused on improving its performance in challenging cases. Considering it as a teaching strategy—rather than giving equal attention to all examples, we focus more on those where the model has difficulty. This is especially useful in our deepfake detection work, where we might have many more examples of one class than the other, and we must ensure that our model learns effectively from both types. To ensure consistent data preprocessing across all datasets, we implemented a comprehensive image augmentation strategy. This included rescaling pixel values, rotating images up to 30 degrees, shifting them both horizontally and vertically by up to 20%, and applying shear transformations, random zooming, horizontal flipping, and subtle brightness adjustments. When required, any missing pixels were filled using the nearest neighbor approach. For the CelebDF-v2 and 140_K real-and-fake faces datasets, we found that processing images in batches of 64 worked the best, with a learning rate of 0.0001. We used the Adam optimizer and Focal Loss function, incorporating both early stopping and learning rate scheduling to prevent overfitting. The models were trained for 50 epochs, although for our CCDDF dataset, we deliberately adjusted the hyperparameter configuration to optimize performance. Unlike the CelebDF-v2 and 140K datasets, CCDDF exhibited better convergence with a larger batch size of 128 and a reduced learning rate of 0.00001. While we maintained the same Adam optimizer, Focal Loss function, and callback mechanisms across all datasets, we found that the CCDDF dataset required only 10 epochs to reach optimal results compared to the 50 epochs needed for the other datasets. Though we conducted experiments using identical hyperparameter settings across all datasets, these dataset-specific adjustments consistently yielded superior performance metrics for CCDDF, highlighting the importance of tailored optimization strategies for different data distributions. Table 3 summarizes the hyperparameters used while training our model.

B. RESULTS

In this subsection, we present and analyze the experimental results of our proposed CNN-MHSA framework across three distinct datasets. We evaluate our model's performance using standard metrics including accuracy, precision, recall, and Area Under the Curve (AUC) scores. The results demonstrate the effectiveness of our sequential integration approach in deepfake detection while maintaining computational efficiency. We also examine the model's attention mechanisms through visualization maps, providing insights into the regions of interest identified during the detection process. Detailed analysis of these results and comparisons with state-of-the-art methods are presented in the following subsections.

TABLE 3. Hyperparameters and their descriptions.

Hyperparameter	140K	Celeb-DFv2	CCDDF
Data Split	71.42% / 27.14% / 1.44%	82.18% / 15.46% / 2.36%	80% / 17% / 3%
Batch Size:	64	64	128
Image Size	224×224×3	224×224×3	224×224×3
Optimizer	Adam	Adam	Adam
Learning Rate	1e-4	1e-4	1e-5
Loss Function	Focal Loss	Focal Loss	Focal Loss
Epochs	50	50	10
Callbacks	EarlyStopping, LR Scheduling	EarlyStopping, LR Scheduling	EarlyStopping, LR Scheduling

1) CCDDF

The experimental results demonstrate excellent performance in our binary classification task. The confusion matrix shows that out of 600 total samples, our model correctly identified 297 fake images and 285 real images. There were only 3 false positives (fake images classified as real) and 15 false negatives (real images classified as fake), resulting in an overall accuracy of 97% The model's strong performance is further validated by the ROC curve analysis, which yielded an impressive Area Under the Curve (AUC) of 0.9958. The ROC curve's shape shows a sharp rise toward the top-left corner of the plot, indicating excellent discrimination ability across different classification thresholds. This is substantially better than a random classifier (shown by the dashed diagonal line), demonstrating that our model has learned meaningful patterns to distinguish between real and fake images. The high AUC score, combined with the low number of misclassifications shown in the confusion matrix, suggests that our model has achieved robust and reliable performance for this deepfake detection task, though there is still a small margin for potential improvement in reducing false negatives. Figure:6 and Figure:7 show the confusion matrix and AUC curve when our proposed model is evaluated on the test set of the generated dataset.

2) 140K REAL AND FAKE FACES DATASET

Our model demonstrated strong performance in distinguishing between real and AI-generated images. The confusion matrix revealed robust classification results across 2000 test samples, with 987 accurate detections of fake images and 967 correct identifications of real images. The model shows relatively low error rates, with only 13 false positives (fake images misclassified as real) and 33 false negatives (real images misclassified as fake), achieving an overall accuracy of 97.7%. The model's exceptional discriminative ability is further validated by an AUC score of 0.9975 on the

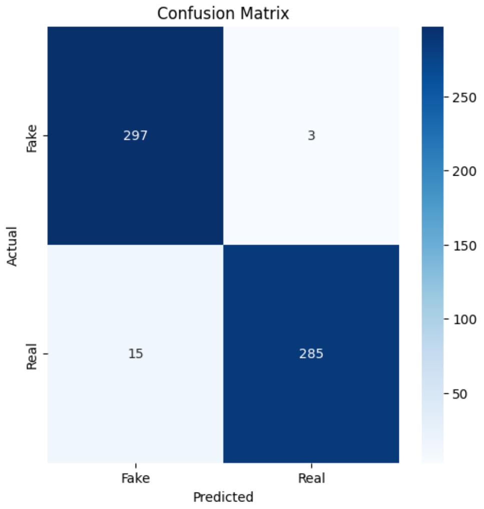


FIGURE 6. Confusion matrix (Own dataset).

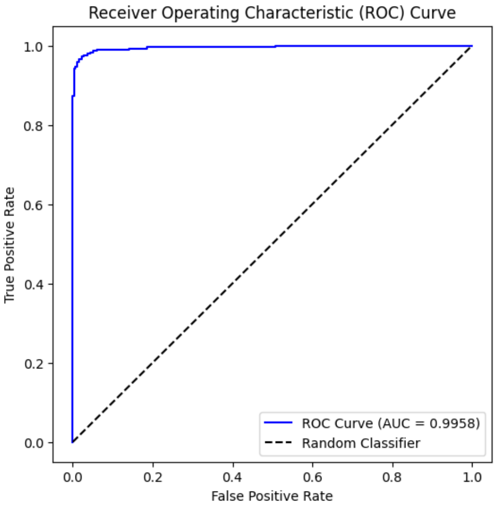


FIGURE 7. ROC curve (Custom dataset).

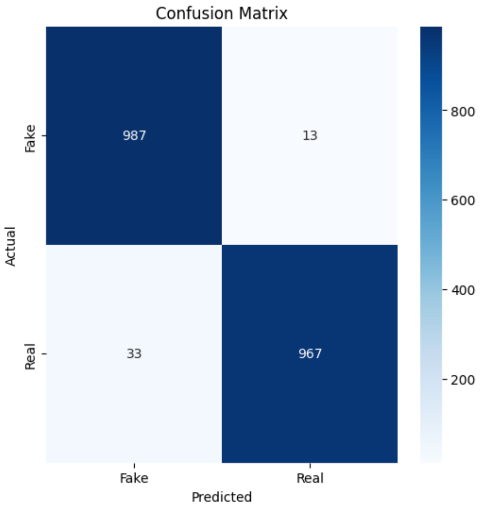


FIGURE 8. Confusion matrix (140k real and fake faces dataset).

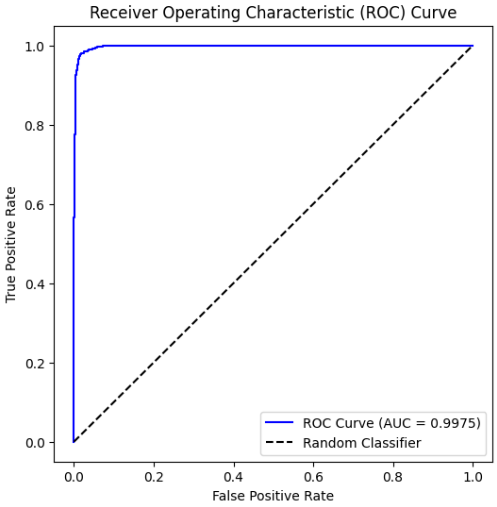


FIGURE 9. ROC curve (140k real and fake faces dataset).

ROC curve. The curve's shape shows a sharp rise to the top-left corner, maintaining high true positive rates while keeping false positive rates minimal. This performance substantially exceeds random classification (indicated by the dashed diagonal line), demonstrating the model's robust ability to distinguish between real and fake images. Figure:8 and Figure:9 show the confusion matrix and AUC curve when our proposed model is evaluated on the test set of the 140k real and fake faces dataset.

3) CelebDF-V2

Our model demonstrated a strong performance in detecting manipulated images. The confusion matrix shows that out of 600 total samples, the model correctly identified 285 fake images and 279 real images. In terms of errors, there were 15 false positives (real images incorrectly classified as fake) and 21 false negatives (fake images missed by the model), resulting in an overall accuracy of 94%. The effectiveness of the model was further validated by ROC curve analysis, which yielded an AUC score of 0.9872. The sharp rise of the curve towards the top-left corner indicates the model's strong ability to maintain a high true positive rate while keeping false positives low. This significantly outperforms random classification (shown by the dashed diagonal line), demonstrating that our model learns meaningful patterns to distinguish between real and manipulated images. Figure:10 and 11 show the confusion matrix and AUC curve when the proposed model was evaluated on the test set of the Celeb-DF v2 dataset.

4) ATTENTION HEATMAP VIZUALIZATION

The attention heatmap visualizations reveal distinct patterns across different classification scenarios. In correctly classified fake images (True: 0, Pred: 0), we observe concentrated attention on facial features that typically exhibit manipulation artifacts. Particularly in the 140K dataset,

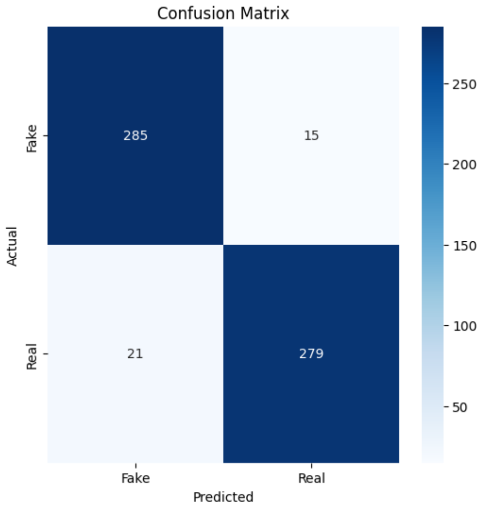


FIGURE 10. Confusion matrix (Celeb DF-v2).

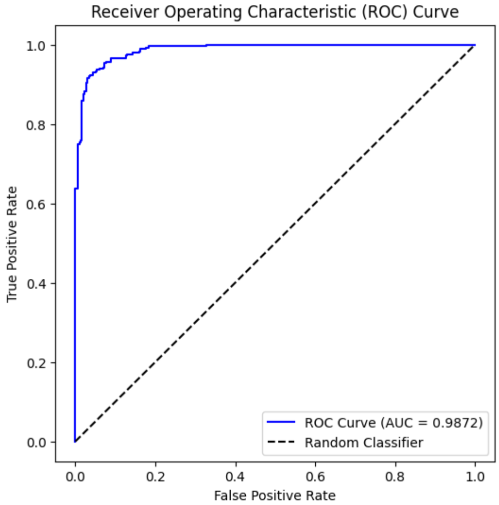


FIGURE 11. ROC curve (Celeb DF-v2).

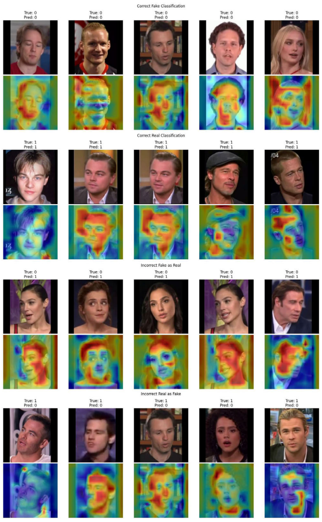


FIGURE 12. Heatmap Visualization of celeb DF V2 dataset.

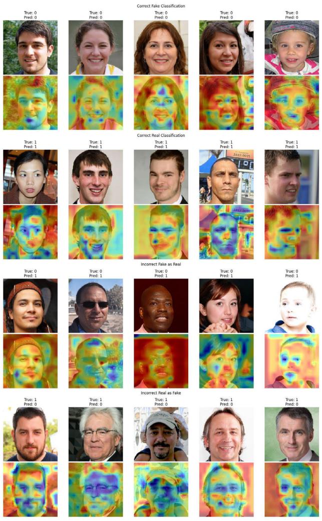


FIGURE 13. Heatmap Visualization of 140k dataset.

Figure 13, attention focuses on facial boundaries, unnatural skin smoothness, and inconsistent lighting patterns that are characteristic of synthetic generation.

For correctly classified real images (True: 1, Pred: 1), attention distributes more evenly across facial features with particular emphasis on eyes, nose, and mouth regions. This balanced pattern is evident across all three datasets, suggesting the model has learned to identify authentic facial characteristics regardless of individual differences.

The misclassification cases provide critical insights into the model's limitations. In fake images incorrectly classified as real (True: 0, Pred: 1), we observe diffused attention or focus on non-discriminative features. For example, in the Celeb-DF v2 dataset, Figure 12, attention often shifts toward background elements rather than facial inconsistencies.

Conversely, in real images incorrectly classified as fake (True: 1, Pred: 0), attention becomes abnormally concentrated on specific facial regions that may resemble synthetic patterns. This is particularly evident in our custom dataset, Figure 14, where the model frequently focuses intensely on eye regions and facial contours in misclassified real images.

Each dataset exhibits distinctive attention patterns reflective of its unique characteristics. The 140K dataset shows more uniform attention on central facial features, the Celeb-DF v2 dataset demonstrates attention influenced by diverse filming conditions, and our custom dataset reveals particularly strong attention to textural details.

C. CROSS-DATASET GENERALIZATION ANALYSIS

To evaluate the real-world robustness of our proposed CNN-MHSA architecture, we conducted rigorous cross-dataset experiments by training models on each dataset and testing their performance on the others. This approach provides critical insights into the model's ability to generalize to unknown deepfake generation techniques.

Table 4 presents the comprehensive results of our cross-dataset evaluation:

TABLE 4. Cross-dataset performance evaluation of deepfake detection models.

Training Dataset	Testing Dataset	Accuracy (%)	AUC Score (%)	Precision (%)	Recall (%)
140K	140K	97.7	99.7	98.7	96.7
140K	Celeb-DF v2	49.7	48.8	49.8	98.3
140K	CCDDF	50.5	60.0	50.3	97.3
Celeb-DF v2	140K	49.6	47.8	49.8	96.8
Celeb-DF v2	Celeb-DF v2	94.0	98.7	94.9	93.0
Celeb-DF v2	CCDDF	52.0	76.8	51.1	97.0
CCDDF	140K	52.8	57.1	51.5	91.5
CCDDF	Celeb-DF v2	49.7	51.4	49.5	32.3
CCDDF	CCDDF	97.0	99.6	99.0	95.0

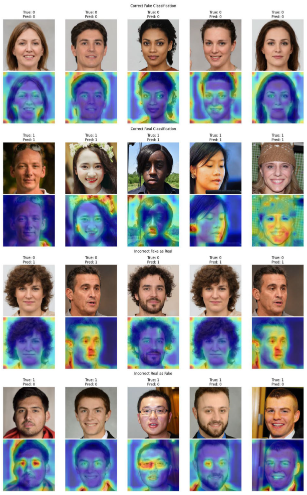


FIGURE 14. Heatmap Visualisation of custom dataset.

The results reveal several important findings. First, our model achieves excellent performance when tested on the same dataset it was trained on, with accuracies ranging from 94.0% to 97.7% and AUC scores between 98.7% and 99.7%. However, there is a significant performance drop in cross-dataset scenarios, with accuracies falling to approximately 50% in most cases. This substantial performance degradation highlights the challenge of cross-dataset generalization. The different deepfake generation techniques used across datasets create distinct artifacts that models learn during training, but these learned patterns may not transfer effectively to other manipulation methods.

To address this limitation, we explored two ensemble approaches:

- **Equal Weights Ensemble:** In this approach, we combined predictions from all three models (trained on 140K, Celeb-DF v2, and CCDDF datasets) with equal weighting (1/3 each) regardless of the test dataset. This simple averaging of predictions yielded noticeable improvements over individual cross-dataset performance, particularly for the Celeb-DF v2 test set where accuracy increased to 77.5% and AUC score to 85.6%. This suggests that even a basic ensemble can capture more generalizable features across different deepfake generation techniques.
- **Dataset-Weighted Ensemble:** This more sophisticated approach assigns weights to each model based on the test dataset being evaluated. Specifically, we assigned a weight of 0.6 to the model trained on the same dataset as the test set and 0.2 to each of the other models. For example, when testing on the 140K dataset, the 140K-trained model received 0.6 weight while Celeb-DF v2 and CCDDF models received 0.2 weight each. This weighted ensemble demonstrated remarkable performance, achieving accuracies between 91.3% and 97.5% across all test datasets, with AUC scores ranging from 97.1% to 99.6%.

Table 5 summarizes the results of the ensemble model performance

The impressive performance of the dataset-weighted ensemble suggests that while individual models may specialize in detecting specific types of manipulation artifacts, combining their predictions allows for more robust detection across diverse generation techniques. This ensemble approach effectively leverages the strengths of each model while compensating for their individual limitations.

These findings highlight an important direction for practical deepfake detection systems: rather than relying on a single model, deploying an ensemble of models trained on diverse datasets can significantly improve real-world performance when encountering unknown manipulation techniques. The weighted ensemble approach in particular shows promise as a strategy for developing more generalizable deepfake detection systems.

D. COMPARISON OF RESULTS

In this section, we present a comprehensive evaluation of the proposed models using the three distinct datasets. First, we showcase the performance metrics of our experimental configurations, highlighting the effectiveness of each model variant under different conditions. Subsequently, we contextualize our findings by comparing them with existing approaches in the literature, demonstrating how our work advances the current understanding in this domain. The comparative analysis not only validates our methodological choices but also illustrates the key advantages and potential limitations of our approach relative to existing solutions. Our comparison studies were strategically selected based on the dataset commonality and methodological diversity, with most utilizing the same 140k dataset to ensure direct benchmarking against established approaches like CNN + SVM [43] and the Custom CNN model [48]. We also included

TABLE 5. Ensemble model performance.

Ensemble Type	Testing Dataset	Accuracy (%)	AUC Score (%)	Precision (%)	Recall (%)
Equal Weights	140K	57.1	96.2	53.9	99.1
Dataset-Weighted	140K	97.5	99.6	95.9	99.4
Equal Weights	Celeb-DF v2	77.5	85.6	70.9	93.3
Dataset-Weighted	Celeb-DF v2	91.3	97.1	88.3	95.3
Equal Weights	CCDDF	57.5	98.2	54.1	99.7
Dataset-Weighted	CCDDF	96.7	99.1	94.9	98.7

TABLE 6. Classification report results from the experiments conducted.

Dataset	Precision		Recall		F1-score		Accuracy (%)
	Fake	Real	Fake	Real	Fake	Real	
CelebDF	0.93	0.95	0.95	0.93	0.94	0.94	94.00
CCDDF	0.95	0.99	0.99	0.95	0.97	0.97	97.00
140k	0.97	0.99	0.99	0.97	0.98	0.98	**98.00**

TABLE 7. Comparison of our results with existing work.

Model	Dataset	Accuracy (%)	AUC Score (%)
CNN + SVM [43]	140k	88.10	88.33
Light-weight CNN [45]	140k	95.0	-
MLP + LSTM [46]	140k	74.7	-
LightFFDNets v2 [47]	140k	71.19	-
Custom CNN model [48]	140k	97.24	-
Defakehop [50]	Celeb-DF v2	-	90.56
Fusion Network [51]	Celeb-DF v2	-	99.1
GAN [57]	Celeb-DF v2	94.51	97.37
GCN + Transformer [58]	Celeb-DF v2	99.47	99.43
CNN + MHSA (Ours)	**140k**	**98.0**	**99.75**
CNN + MHSA (Ours)	Celeb-DF v2	94.0	98.72
CNN + MHSA (Ours)	**CCDDF**	**97.0**	**99.58**

recent high-performers such as the Fusion Network [51] on the Celeb-DF v2 dataset to position our work within cutting-edge research.

It is important to note that while our model demonstrates competitive performance, we identified two recent approaches that have achieved superior results specifically on the Celeb-DF v2 dataset [56]. Li et al. [57] achieved a remarkable accuracy of 94.51% using their innovative GAN-based detection framework on the Celeb-DF v2 dataset, slightly outperforming our 94.0% accuracy on the same dataset. More significantly, Khormali and Yuan [58] reported an impressive 99.47% accuracy on Celeb-DF v2 by leveraging a sophisticated GCN + Transformer architecture, which represents the current state-of-the-art performance on this specific dataset.

The results presented in Table 7 reveal important insights about model performance across different contexts. While our base CNN + MHSA model achieves strong results when trained and tested on the same dataset (97.7% accuracy on 140k, 94.0% on Celeb-DF v2, and 97.0% on CCDDF), our cross-dataset evaluation in Table 4 exposes a critical challenge in the field: most models, including ours, suffer dramatic performance degradation when tested on datasets different from their training data. For instance, our model trained on 140k achieves only 49.7% accuracy when tested on Celeb-DF v2, revealing the substantial domain gap between these datasets. This limitation motivated our development of an ensemble approach, with results presented in Table 5. Our dataset-weighted ensemble strategy significantly improves cross-dataset generalization, achieving 97.5% accuracy on 140k, 91.3% on Celeb-DF v2, and 96.7% on CCDDF. This represents a substantial improvement over both our base model and competing approaches when considering performance across multiple datasets simultaneously. The MHSA mechanism remains a key strength of our architecture, enabling effective capture of spatial relationships and manipulation artifacts that often elude traditional convolutional approaches. This advantage is particularly evident in the dataset-weighted ensemble's balanced performance across all evaluation metrics - accuracy, AUC, precision, and recall - demonstrating robust detection capabilities across varied deepfake generation techniques and image characteristics.

V. CONCLUSION

This study presents an experimental approach to deepfake detection by combining multi-head self-attention mechanisms with convolutional neural networks. The proposed model demonstrated remarkable effectiveness across multiple datasets, achieving high accuracy and AUC scores that validated its robustness and reliability. Specifically, the model achieved 98% accuracy with an AUC of 99.75 on the 140K Real and Fake Faces dataset, 94% accuracy with an AUC of 98.72 on the Celeb-DF v2 dataset, and 97% accuracy with an AUC of 99.58 on our custom-generated dataset. These results significantly outperform many existing approaches and demonstrate the model's ability to generalize across different types of deepfake content. We recognize that our evaluation of the Celeb-DF v2 dataset would be strengthened by explicitly testing across different compression levels, as deepfake content in real-world scenarios often undergoes compression when shared on social media platforms. While our current model has shown robust performance on the standard Celeb-DF v2 dataset, future work will specifically evaluate degraded image quality scenarios and incorporate compression-aware training strategies to enhance real-world applicability. The integration of attention heatmap visualizations has provided valuable insights into the model's decision-making process, enhancing our approach's interpretability and trustworthiness. This transparency is crucial for the practical deployment of deepfake detection systems in real-world scenarios. The strong performance across diverse datasets suggests that our model is robust to various deepfake generation techniques and can adapt to new forms of synthetic media. This study demonstrates that an architecture combining CNN and MHSA mechanisms can achieve effective deepfake detection across diverse datasets. With 696,001 parameters, our model provides strong detection performance on multiple benchmark datasets, contributing to the ongoing research in identifying manipulated media content. Future work could focus on expanding the model's capabilities to handle video content, exploring real-time detection capabilities, and investigating the model's resilience to newly emerging deepfake generation technologies. Additionally, developing lightweight versions of the model could facilitate deployment on resource-constrained devices, making reliable deepfake detection more accessible to end-users. As deepfake technology continues to evolve, our work provides a strong foundation for developing increasingly sophisticated detection methods to help maintain trust in digital media.

REFERENCES

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, Jan. 2021.

[2] D. Johnson, T. Gwyn, L. Qingge, and K. Roy, "Deepfake detection using CNN trained on eye region," in *Proc. Int. Conf. Ind.*, Jan. 2022, pp. 443–451.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 73–76.

[4] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023.

[5] N. Caporusso, "Deepfakes for the good: A beneficial application of contentious artificial intelligence technology," in *Proc. Adv. Artif. Intell.*, Jul. 2020, pp. 235–241.

[6] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Comput. Vis. Image Understand.*, vol. 223, Oct. 2022, Art. no. 103525.

[7] (2024). *Deepfakes of Your Dead Loved Ones Are a Booming Chinese Business*. Accessed: Oct. 5, 2024. [Online]. Available: https://www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/

[8] (2021). *Deepfake Lip Syncing Technology Could Help Translate Film and Tv Without Losing an Actor's Original Performance*. Accessed: Oct. 5, 2024. [Online]. Available: https://www..be/deepfake-lip-syncing-technology-could-help-translate-film-and-tv-without-losing-an-actors-original-performance/

[9] (2020). *Ai Deepfake Memes*. Accessed: Oct. 7, 2024. [Online]. Available: https://www.technologyreview.com/2020/08/28/1007746/ai-deepfakes-memes/

[10] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[12] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 83–92.

[13] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.

[14] M. Bonomi, C. Pasquini, and G. Boato, "Dynamic texture analysis for detecting fake faces in video sequences," *J. Vis. Commun. Image Represent.*, vol. 79, Aug. 2021, Art. no. 103239.

[15] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.

[16] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.

[17] M. A. S. Habeeba, A. Lijiya, and A. M. Chacko, "Detection of deepfakes using visual artifacts and neural network classifier," in *Innovations in Electrical and Electronic Engineering*. Cham, Switzerland: Springer: Jul. 2020, pp. 411–422.

[18] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jan. 2019, pp. 5001–5010.

[19] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on the discrepancy between the face and its context," *arXiv:2008.12262*, 2020.

[20] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2019, pp. 1–6. [Online]. Available: https://api.semanticscholar.org/CorpusID:196622700

[21] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1053–1061.

[22] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other- audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 439–447, doi: 10.1145/3394171.3413700.

[23] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao, "DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms," in *Proc. 28th ACM Int. Conf. Multimedia*, vol. 28, Jan. 2020, pp. 1318–1327.

[24] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2018, pp. 1–7.

[25] A. Rössler, D. Cozzolino, L. Verdoliva, C. Rieß, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1–11.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[27] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[28] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," 2018, *arXiv:1806.02877*.

[29] I. Ganiyusufoglu, L. M. Ngô, N. Savov, S. Karaoglu, and T. Gevers, "Spatio-temporal features for generalized detection of deepfake videos," 2020, *arXiv:2010.11844*.

[30] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 053–058.

[31] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2019, pp. 2307–2311.

[32] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," 2019, *arXiv:1910.12467*.

[33] N. Bonettini, E. Daniele Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," 2020, *arXiv:2004.07676*.

[34] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[35] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.

[36] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable deepfake detection with locality-aware AutoEncoder," 2019, *arXiv:1909.05999*.

[37] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometrics Theory*, Sep. 2019, p. 1, doi: 10.1109/BTAS46853.2019.9185974.

[38] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions Don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2823–2832, doi: 10.1145/3394171.3413570.

[39] A. Khormali and J.-S. Yuan, "ADD: Attention-based DeepFake detection approach," *Big Data Cognit. Comput.*, vol. 5, no. 4, p. 49, Sep. 2021.

[40] S. Dasgupta, J. Mason, X. Yuan, O. Odeyomi, and K. Roy, "Enhancing deepfake detection using SE block attention with CNN," in *Proc. Int. Conf. Artif. Intell.*, Aug. 2024, pp. 1–6.

[41] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to Spot…for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8695–8704.

[42] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3207–3216.

[43] J. Mallet, L. Pryor, R. Dave, and M. Vanamala, "Deepfake detection analyzing hybrid dataset utilizing CNN and SVM," in *Proc. 7th Int. Conf. Intell. Syst.*, Apr. 2023, pp. 7–11.

[44] (2020). *140k Real and Fake Faces*. Accessed: Feb. 17, 2024. [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

[45] T. Arshad, M. H. Khan, and M. S. Farid, "An efficient framework to recognize deepfake faces using a light-weight CNN," in *Proc. 9th Int. Conf. Multimedia Syst. Signal Process. (ICMSSP)*, May 2024, pp. 24–29.

[46] J. Mallet, N. Krueger, R. Dave, and M. Vanamala, "Hybrid deepfake detection utilizing MLP and LSTM," in *Proc. 3rd Int. Conf. Electr.*, Jul. 2023, pp. 1–5.

[47] G. Jabbarlı and M. Kurt, "LightFFDNets: Lightweight convolutional neural networks for rapid facial forgery detection," 2024, *arXiv:2411.11826*.

[48] S. Pawar, S. Vetal, S. Sharma, A. Pagare, and C. Patil, "Generation and detection of artificial facial images," in *Proc. 8th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Aug. 2024, pp. 1–6.

[49] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5781–5790.

[50] H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, and C.-C.-J. Kuo, "DefakeHop: A light-weight high-performance deepfake detector," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2021, pp. 1–6.

[51] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez, "DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation," *Eng. Appl. Artif. Intell.*, vol. 110, Apr. 2022, Art. no. 104673.

[52] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2018, *arXiv:1812.04948*.

[53] S. Dasgupta and K. Badal, 2025, "Center-for-cyber-defense-deepfake-dataset," doi: 10.21227/ddmc-5g05.

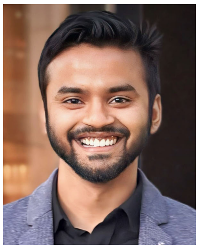
[54] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[56] P. Edwards, J.-C. Nebel, D. Greenhill, and X. Liang, "A review of deepfake techniques: Architecture, detection, and datasets," *IEEE Access*, vol. 12, pp. 154718–154742, 2024.

[57] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1658–1670, Apr. 2023.

[58] A. Khormali and J.-S. Yuan, "Self-supervised graph transformer for deepfake detection," *IEEE Access*, vol. 12, pp. 58114–58127, 2024.



SUBHRAM DASGUPTA was born in Kolkata, India, in October 1997. He received the B.Tech. degree in computer science from Christ University, Bengaluru, India, in 2019, and the M.S. degree in computer science from North Carolina A&T State University, in 2024. He is currently pursuing the Ph.D. degree with North Carolina A&T State University, with a focus on cybersecurity. He is a Teaching Assistant while pursuing the Ph.D. degree. He is also a member with the Center of Cyber Defense (CCD), Cyber Defense and AI Laboratory. Throughout his academic career, he has developed expertise in computational analysis and innovative research approaches, while demonstrating exceptional collaborative abilities and dedication to expanding his knowledge in emerging technologies. He aims to leverage his technical expertise and innovative problem-solving approaches to make significant contributions to the fields of cybersecurity and artificial intelligence. His research interests include machine learning algorithms, deep learning frameworks, computer vision technologies, and large language models.

KUSHAL BADAL received the master's degree in computer science from North Carolina A&T State University, with a strong academic record and a passion for AI and data science. He gained practical experience in data analysis, machine learning, and research, complemented by leadership roles and a commitment to continuous learning. His goal is to leverage his technical skills and enthusiasm for problem-solving to make meaningful contributions in the field of artificial intelligence. His research interests include cyber security, machine learning, deep learning, computer vision, large language model, and problem solving.

SWETHA CHITTAM received the bachelor's degree in electrical and electronics engineering from Osmania University, India, and the master's degree in computer science from North Carolina A&T State University, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. She is a member with the Center of Cyber Defense (CCD), Cyber Defense and AI Laboratory. She specializes in artificial intelligence (AI) and data science. Her research interests include AI, machine learning(ML), deep learning (DL), computer vision, biometrics, data science, big data analytics, cybersecurity, cyber-physical systems, and explainable AI (XAI). Her current research focuses on deepfake detection and generation using deep learning models and generative adversarial networks (GANs).

MD TASNIM ALAM received the Bachelor of Technology degree in computer science and engineering from NIT Rourkela, Odisha, India. He is currently pursuing the master's degree with the Department of Computer Science, North Carolina A&T State University, conducting research at the Cyber Defense and AI Laboratory. Prior to his graduate studies, he was a Lecturer with the Department of Computer Science, North Bengal International University. At NC A&T, he is a Graduate Research and Teaching Assistant. His work focuses on computer vision, machine learning, deep learning, artificial intelligence (AI), and explainable AI (XAI), with a particular emphasis on human action recognition (HAR). He has published papers on deep learning, AI, XAI, and computer vision at international conferences, focusing on optimizing human action recognition using deep learning models and XAI tools. He is also involved in a National Science Foundation (NSF) funded project.

KAUSHIK ROY (Senior Member, IEEE) is currently a Professor and the Chair of the Department of Computer Science, North Carolina A&T State University (NCAT). He holds the Jefferson-Pilot/Ron McNair endowed chair position at the Department of Computer Science. He is the Director of the Center for Cyber Defense (CCD). He directs the Center for Trustworthy AI and the Cyber Defense and AI Laboratory. He has over 210 publications, including 61 journal articles and a book. His research is funded by the National Science Foundation (NSF), Department of Defense (DoD), and the Department of Energy (DoE) and Boeing. His current research interests include cybersecurity, the IoT, cyber-physical systems, cyber identity, biometrics, AI and machine learning, data science, and big data analytics. He is an Associate Editor of IEEE SECURITY AND PRIVACY.