Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Attention-Based VGG19-ResNet50 Ensemble for Generalizable Deepfake Image Detection- To Mitigate Cyberbullying

Swetha Chittam

schittam@aggies.ncat.edu

North Carolina Agricultural and Technical State University

Subhram Dasgupta

North Carolina Agricultural and Technical State University

Sophina Luitel

North Carolina Agricultural and Technical State University

Madhuri Siddula

North Carolina Agricultural and Technical State University

Kaushik Roy

North Carolina Agricultural and Technical State University

Research Article

Keywords: Ensemble model, VGG19, ResNet50, Squeeze-And-Excitation Block, SE Block, Deepfake Image detection, CCD-FusedFaces, 140k-Real-And-Fake Faces, Cyber Bullying, Child Abuse

Posted Date: June 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9707209/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Attention-Based VGG19-ResNet50 Ensemble for Generalizable Deepfake Image Detection- To Mitigate Cyberbullying

Swetha Chittam[1*], Subhram Dasgupta[1†], Sophina Luitel[1†], Madhuri Siddula[1†], Kaushik Roy[1†]

[1]Department of Computer Science, North Carolina Agricultural and Technical State University, East Market Street, Greensboro, 27411, North Carolina, USA.

*Corresponding author(s). E-mail(s): schittam@aggies.ncat.edu;
Contributing authors: sdasgupta@aggies.ncat.edu; sluitel@aggies.ncat.edu; msiddula@ncat.edu; kroy@ncat.edu;
†These authors contributed equally to this work.

Abstract

With advanced image generation techniques and rapidly evolving deep learning technology, deepfakes have become a significant concern, with increasingly realistic fake content on social media impacting people's lives. Deepfakes raise serious identity and trust issues, as they are frequently used to spread misinformation, manipulate public opinion, damage reputations, and most concerning child abuse via cyberbullying. What makes this threat even more alarming is that such fabricated content is not only relatively easy to produce but also extremely difficult to detect. To address this issue, we developed a robust ensemble method with and without a Squeeze-and-Excitation (SE) block to detect real vs fake images. In this study, we employed pretrained convolutional neural networks (CNNs), specifically VGG19 and ResNet50, in our ensemble model. To further enhance the model, we incorporated a channel-wise attention mechanism (SE block) that recalibrates channel-wise features, and allows the network to focus on the most informative feature maps. The two models we developed are VGG19-ResNet50 and VGG19-ResNet50-SE. We trained and evaluated these models on our own CCD-FusedFaces dataset and to evaluate generalization we tested these models on 140k-Real-And-Fake Faces. Additionally, we performed domain-enhanced training to improve the models' generalization capability. Our results

demonstrated that our models achieved exceptional performance, with VGG19-ResNet50-SE (Fusion Model with Attention) achieving 99.50% test accuracy and VGG19- ResNet50 with Partial Fine-Tuning achieving 99.41% test accuracy on the CCD-FusedFaces dataset, significantly outperforming state-of-the-art methods.

Keywords: Ensemble model, VGG19, ResNet50, Squeeze-And-Excitation Block, SE Block, Deepfake Image detection, CCD-FusedFaces, 140k-Real-And-Fake Faces, Cyber Bullying, Child Abuse

1 Introduction

Deepfakes are becoming increasingly prevalent across social media and other digital platforms. These manipulated media can take various forms, including audio, images, and videos. While the technology behind deepfakes has potential for creative and educational applications, its misuse poses serious threats; especially in sensitive domains such as law enforcement, identity theft, cyberbullying, misinformation campaigns, defamation of public figures, and, most alarmingly, child abuse via cyberbullying. This work primarily focuses on identifying and protecting children from threats like online child abuse and cyberbullying in the digital age.

Deepfake technology is increasingly used to create non-consensual, AI-generated child sexual abuse material (CSAM), contributing to digital forms of child abuse. The malicious actors use deepfakes to fabricate realistic pornographic content involving children, often using altered images or likenesses of real minors. This causes severe psychological and reputational harm to victims [1, 2]. The advancements in digital technology, especially generative AI and deepfakes, are enabling the creation and distribution of highly realistic, AI-generated child sexual abuse material (CSAM), posing a grave threat to child safety. Offenders exploit open-source AI tools to produce synthetic images that mimic real child abuse by often sharing techniques and models on dark web forums to bypass content moderation and legal restrictions. The digital deepfake technology is not only fueling harmful fantasies but also expanding the ecosystem of child abuse, demanding urgent legal and policy responses worldwide [2]. Additionally, with the widespread popularity of online games, particularly among younger children and teenagers, there is growing concern. Many children are naturally outgoing, curious and sociable, but they lack the maturity and awareness needed to understand the risks of digital interactions. As a result, they often interact with strangers online and engage in multiplayer games without understanding the potential consequences. During interaction over online games or other social media platforms, children may unknowingly share personal information through in-game chats, or they might participate in video calls or voice chats, sharing personal information or photos. They are often unaware that not everyone they interact with online is trustworthy. Malicious individuals such as predators or cyberbullies, may target these children, record their interactions, and exploit the shared content. Such attackers can use this personal data to create fake images, voice clips, or videos through deep fake technology. They may use fake audio

or video impersonating a child's parents to deceive and manipulate them into taking harmful actions. Manipulated content is increasingly being used to threaten or blackmail children and attackers often claiming they will share the fake content with the child's school or peers. Fearing public embarrassment and humiliation, the child may feel pressured to comply with the demands. This can lead to serious emotional consequences, including anxiety, depression, and mental-health issues. In more severe cases, abusers may use deepfakes to blackmail a child's parents, escalating their threats by demanding money through fabricated and manipulative content. The real-world examples, such as a case in North Carolina where a child psychiatrist used AI to convert school photos of children into explicit content and another in Quebec where a man was sentenced for producing synthetic child pornography using deepfake tools and images from social media. The study emphasises how digital exploitation, even without AI can lead to severe psychological distress and suicide. These cases illustrate how deepfakes intensify existing threats, making it easier to fabricate realistic abuse content and target children without their knowledge or consent [3]. Although parents play a vital role in protecting their children through parental controls, supervision, and education. The responsibility cannot rest on families alone. As researchers and technologists, we share the duty to address these challenges by developing robust deep fake detection tools by enhancing child-centered online safety measures, and raising public awareness. Since, deepfakes are increasingly being used to exploit vulnerable individuals, especially children. It is imperative that, we create effective solutions to detect and prevent such threats to ensure a safer digital environment. Additionally, urgent global action is needed through legal reforms, public education, and stronger content moderation by platforms to address this alarming form of digital child abuse.

To address this issue, we propose a generalizable and efficient deep learning ensemble model and to test the robustness of the model, we aim to generate a custom synthetic dataset named Center of Cyber Defense (CCD)-FusedFaces using an open-source AI tool called Face Fusion [4]. The motivation behind generating our own dataset is to contribute to the research community by providing diverse and realistic samples, that can support the development and evaluation of deepfake detection models across various manipulation techniques. In the context of addressing cyberbullying and child abuse cases, the dataset will not only include images of male and female indviduals across different age groups such as younger and older adults but also images of children specifically school-aged children and teenagers, thus allows us to test the model's effectiveness in scenarios involving minors. In this study, we employed a combination of two pretrained Convolutional Neural Networks (CNN) models namely Visual Geometry Group (VGG19, a lightweight model) and a Residual Network (ResNet-50, a heavier model). The motivation behind choosing VGG19 lies in its uniform architecture and low computational cost. However, to address the trade-off between accuracy and efficiency, we incorporated ResNet-50, a powerful architecture known for extracting deep semantic features in image classification tasks. Although VGG19 and ResNet50 have been used individually or in combination with other models for image classification and deepfake detection tasks, to the best of our knowledge, there is no prior work that explicitly explores an ensemble model combining VGG19

and ResNet50. Moreover, numerous ensemble models exist, but they lack model generalization capabilities[5] [6]. To address this gap, we developed an ensemble model by integrating both architectures, with and without the inclusion of a Squeeze-and-Excitation (SE) block, using domain-enhanced training. This allowed us to investigate how effectively these pretrained models can be leveraged to construct a more efficient and robust ensemble for deepfake detection. VGG19 is chosen for its computational efficiency, requiring less processing power and offering faster inference times, making it ideal for systems with limited hardware or real-time applications. In recent studies, VGG19 has consistently achieved high accuracy of 95% for deepfake detection tasks. Its pretrained availability on ImageNet dataset facilitates efficient fine-tuning for transfer learning. Moreover, VGG19 has demonstrated strong generalizability across diverse deepfake datasets [7][8][9][10]. On the other hand, ResNet-50 was chosen for this study, as it is a popular deep learning model that effectively addresses the vanishing gradient problem through its residual connections. The model also ensures efficient gradient flow by enabling deep architectures to learn complex patterns. This structure allows ResNet-50, to capture both high-resolution visual features and essential semantic cues, making it particularly effective in identifying the subtle facial manipulations typical of deepfake imagery. Its depth and robust feature learning make it adept at extracting nuanced expressions and inconsistencies in synthetic faces. Prior studies [11] [12] [13] have also highlighted its strength in deepfake detection.

In this study, VGG19 and ResNet50 were used as baseline models and served as feature extractors. We evaluated their performance individually as well as in combination(VGG19-ResNet50) called Fusion Model, both with and without the integration of Squeeze-and-Excitation (SE) blocks. A custom classifier (MLP) was employed for binary classification purposes. The models were trained and tested on two datasets: our own custom dataset,CCD-FusedFaces and an open source dataset, 140k-Real-and-Fake-Faces[14]. This ensemble model architecture was designed to assess the impact of SE blocks on the model performance, comparing configurations with and without SE integration. Our experiments were conducted using the PyTorch framework. For learning rate tuning, we employed the ReduceLROnPlateau strategy. We also focused on improving the model generalization by applying domain-enhanced training, in which a new training set is constructed by combining datasets from multiple domains. This approach demonstrated that, when deep learning models are exposed to datasets from varying domains, their performance becomes more robust and applicable to real-world scenarios. Although the Fusion Model with Attention achieved higher accuracy in our study, the Fusion Model without Attention demonstrated more consistent performance across all datasets and is computationally efficient. The two models developed in this study are tested on datasets generated using different methods. We tested these models using our own custom dataset named CCD-FusedFaces. We also used the 140k-Real-And-Fake Faces dataset, which contains images generated using StyleGAN.This dataset provides rich visual variability, with faces appearing across a broad spectrum of lighting conditions, including indoor and outdoor settings, changes in natural light, and a variety of background scenes. The CCD-FusedFaces dataset consists of 60,000 real images from Flickr-faces-FFHQ [15] dataset and 20,000 AI-generated fake images for Face Swap and Expression use cases for all demographic

information such as Men, Women and Children, while the 140k-Real-And-Fake-Faces dataset [14] includes both real and StyleGAN-generated images of 70k images for each class. Our models demonstrated consistent performance across both datasets and showed improved results when domain-enhanced training was applied.

The key contributions of this study are:

- Creation of the CCD-FusedFaces dataset.
- The design and evaluation of the ensemble model with and without the integration of Squeeze-and-Excitation (SE) blocks, using an ablation technique.
- Implementation of a domain-enhanced training strategy to improve model generalisation.

The rest of this paper is organised as follows. Section 2 provides a literature review of the CNN models: VGG19, ResNet50, existing Ensemble models,SE block mechanism and general deep fake detection approaches. Section 3 provides Methodology, Data collection, augmentation, preprocessing, building ensemble model and classification. Section 4 provides the results of the study and comparative analysis with existing state-of-the-art methods. Finally, Section 5 concludes with implications, limitations, and suggestions for future research.

2 Related Works

This section explores the technologies and methodologies used in recent studies, that underpin our ensemble-based deepfake detection framework. We begin by examining VGG-19's pioneering approach to deep convolutional architectures and its impact on feature extraction capabilities. We then investigate, ResNet-50's residual learning mechanism that enabled training of deeper networks while maintaining gradient flow. The discussion progresses to ensemble methods, highlighting how combining multiple models can overcome individual limitations and achieve superior classification performance. We analyze the Squeeze-and-Excitation block and channel-wise attention mechanisms that allow networks to adaptively emphasize the most relevant features for detection tasks. Finally, we review existing deepfake detection approaches to understand current challenges and identify opportunities for improvement that our proposed methodology addresses.

2.1 Visual Geometry Group (VGG19):

In the study by Pashine et. al. [7], VGG-19 was trained on deep fake image data and achieved good accuracy (around 73% on the test set) with low errors. While not the most accurate compared to more complex models like Xception, VGG-19 stands out for its computational efficiency, it requires less processing power and memory, and has much faster inference time, especially when optimized with TensorRT (just 113.9 milliseconds). This makes VGG-19 a great choice for systems with limited hardware, where speed and efficiency are important for real-time applications. Hayathunnisa et al. [8] used VGG19 as one of the models to classify real versus AI-generated images from the CIFAKE dataset. VGG19 achieved a validation accuracy of 84.85% after 20 training epochs. While it performed well and completed training in just 44 minutes, it

was outperformed by a custom-designed CNN model that achieved 94.46% accuracy, but it took more than 3 hours for model training. Therefore, despite its slightly lower accuracy, VGG19 was highlighted for its computational efficiency and effectiveness in feature extraction and demonstrates high performance in identifying fake images. VGG19 is employed as the backbone model in the wavelet-integrated deep learning model for deepfake detection.This study highlights VGG19's ability to extract high-level abstract features through a simple and uniform architecture composed of 19 layers. VGG19's architecture consisting of 3×3 convolutional filters and ReLU activations, along with its deep and wide structure, makes it effective for identifying subtle manipulation artifacts in facial images. Its pretrained availability on ImageNet allows for efficient fine-tuning, enabling strong performance in transfer learning tasks[10]. In this study, the wavelet-enhanced VGG19 model achieved robust results across multiple deepfake datasets, and outperforms more complex architectures like Xception in cross-domain evaluations and thus demonstrating superior generalizability and effectiveness in detecting fake faces. Taeb et al. [9] evaluated VGG19 as part of a comparative study on deepfake detection techniques and found it to be highly effective and observed that the model achieved 95% accuracy, the highest among the models tested. The authors stated that VGG19 can extract rich and discriminative features from facial images and makes it well-suited for tasks such as face recognition and deepfake classification. The study also highlighted VGG19's effectiveness in both full-face and partial-face classification settings.

2.2 ResNet50

ResNet-50 was evaluated as one of the four deep learning models for deepfake detection[16]. ResNet-50, a 50-layer deep residual network, effectively addresses the vanishing gradient problem through residual connections and enables gradient flow across layers. This architecture allows the model to learn, both high-resolution features in early layers and importance maps in later layers, and enhances its capacity to distinguish subtle deepfake manipulations. ResNet-50 proved to be a stable and accurate model for detecting deepfakes, showing promise in complex computer vision tasks. Niteesh Kumar et al. [13] used pretrained models-VGG16, ResNet50, and InceptionV3, individually for feature extraction from deepfake images, with each model's features passed separately to classifiers such as SVM, Logistic Regression, and Random Forest. The features from different models are not combined; instead, each model-classifier pair is evaluated independently to assess performance. The authors conclude that the combination of ResNet50 for feature extraction and Logistic Regression for the classification task, delivers the highest accuracy of 88.47% in detecting deepfake images. Lew Kar Yee et al. [17] presents a deep fake detection approach using the ResNet50 model combined with Gaussian blur as a pre-processing technique to improve generalization. The study uses two datasets namely DeepFakeFace and Deepfake Processed, and evaluates performance with different train-test splits (60:20:20 and 80:10:10). ResNet50 is fine-tuned with additional layers and trained on 4,000 images per experiment. Gaussian blur is applied with varying kernel sizes to suppress low-level artifacts. The best accuracy achieved is 82.75% without blur, while blur improves specificity up to 65%, reducing false positives. The results show that Gaussian blur enhances model

robustness and makes ResNet50 more effective for detecting deepfakes across diverse datasets. Pranav Sathe et al. [18] evaluates and compares the performance of several deep learning models, including CNN, CNN+LSTM, ResNet50, and YOLOv8, for detecting deepfake images. Among these, ResNet50 is notable, for its strong feature extraction capabilities which leverages its deep architecture and residual connections to effectively learn subtle differences between real and fake images. The model achieved 74% accuracy, 75% precision, and a 73% F1-score, outperforming both standard CNN and CNN+LSTM architectures.

The authors highlight that ResNet50's depth and residual design enable it to capture complex facial features often missed by shallower models, making it a powerful tool for DeepFake detection. While more computationally demanding, ResNet50 provides a solid balance between accuracy and robustness. Its performance in the study underscores, its value as a reliable and effective architecture for detecting manipulated facial content. Soha Safwat et al. [12] introduces a hybrid model combining ResNet50 and GAN to enhance deepfake detection. ResNet50 serves as a powerful feature extractor, strengthened with Channel-Wise Attention Mechanisms to focus on key facial features and manipulation artifacts. Its deep architecture and residual connections helps in efficient training process and robust feature learning. While ResNet50 alone outperforms models like VGG16, the hybrid model achieves even better performance, with over 83% accuracy and shows improved generalization and precision. This highlights ResNet50's effectiveness and scalability in detecting fake faces. Yesugade, Kiran, et al. [11] states that ResNet-50 demonstrates a strong capacity for precise image analysis. This study states that, ResNet50 has shown higher precision in classifying metastatic cancer images. Its robustness in classification tasks, particularly with complex datasets, is also a notable advantage. Furthermore, ResNet-50's ability to capture subtle features and effectively classify facial expressions makes it highly suitable for deepfake detection.

2.3 Ensemble models

In many recent studies (Patel, Yogesh, et al. [19], Ganguly, Shreyan et al. [20], Tay, Wan Xin, et al. [21]), ensemble models are used for deep fake image or video detection tasks and have shown promising results. The features extracted from two different deep learning models are stacked together and then passed to the classifier. The importance of ensemble models lies in their ability to overcome the limitations of single CNN architectures.In general single models, often struggle to generalize across different types of manipulations or datasets. By incorporating attention mechanisms, as an ensemble enhancement, the model gains the capacity to detect a wider range of forgery artifacts, such as inconsistent textures, edges, or facial features. In this study by Ganguly, Shreyan, et al. [20], ViXNet, an ensemble deepfake detection model is employed that integrates a Vision Transformer (ViT) with a Xception CNN to effectively capture both local and global facial features indicative of manipulation. A key innovation in ViXNet is, use of patch-wise self-attention, where an input image is divided into patches that are selectively masked using learnable convolutional filters. This enables the model to emphasize regions likely to contain deepfake artifacts. These masked patches are processed by ViT, whose multi-head self-attention mechanism captures

long-range dependencies and global correlations between distant facial regions and thus allows ViXNet to identify subtle and spatially dispersed inconsistencies often missed by standard CNNs. The features extracted from the ViT and Xception branches are fused for classification, enabling robust detection of forgeries. ViXNet demonstrates strong performance on benchmark datasets such as FaceForensics++, Celeb-DF (V2), DFID, and DFDC, achieving up to 98.93% AUC in intra-dataset settings and showing excellent generalization with up to 83.60% AUC in inter-dataset tests, outperforming several state-of-the-art models. A novel deepfake detection architecture called RSE-Net, which combines two variants of the pretrained EfficientNetB0 model. One variant (Model A) is fine-tuned by replacing the original Swish activation with ReLU to reduce overfitting and improve efficiency, while the other (Model B) is frozen to preserve generic features learned from ImageNet. The outputs from both models are fused and passed through additional dense and softmax layers for classification. This ensemble approach leverages the strengths of both feature adaptation and generalization to robustly detect manipulated videos across multiple datasets. EfficientNetB0 is used not only for fixed feature extraction but also for fine-tuning in one stream, enabling the model to learn deepfake-specific patterns. Its role in the ensemble improves performance, efficiency, and generalizability across different types of deepfakes. The model achieves state-of-the-art results when tested against FaceForensics++ and CelebDF datasets, and demonstrates that integrating pretrained CNNs in both frozen and trainable forms can significantly enhance detection accuracy and robustness (Minhas, Samiya Afzal et al., [22]). Inspired by these studies (Patel, Yogesh, et al. [19], Ganguly, Shreyan et al. [20], Tay, Wan Xin, et al.[21], Minhas, Samiya Afzal et al.[22]), we developed an ensemble model by combining VGG19 and ResNet50 for feature extraction. In this ensemble model ResNet50 is not just used for feature extraction but also finetuned, which enables the model to learn specific patterns from input data and enhances model performance. The features from VGG19 and ResNet50 are stacked together and passed to the custom classifier for binary classification. Ha, Hyunsoo, et al (2023) proposed hybrid models combining vision transformers and CNNs, which have shown strong performance in DeepFake detection, particularly for side faces and low-quality videos. A weighted majority voting ensemble of ResNeSt269 and DeiT achieved 97.66% accuracy on the DFDC dataset, surpassing the previous state-of-the-art accuracy of 96.78%[23]. Another recent study demonstrated that combining CNNs with Vision Transformers (ViTs) enhances deepfake detection performance. The proposed hybrid model, MesoNet-ViT, achieved higher accuracy than standalone methods, with MesoNet-ViT reaching 96.2% on DFDC and 98.1% on DF-TIMIT (Ismail Ilhan et al, 2025)[24].

2.4 Squeeze-And-Excitation Block (SE Block)

Squeeze-and-Excitation (SE) blocks have emerged as a transformative mechanism for enhancing deepfake detection through adaptive channel-wise attention. Originally introduced by Hu et al. [25], SE blocks have capability to address a critical limitation in traditional convolution operations by explicitly modeling interdependencies between

feature channels rather than treating all channels equally. This ability proves invaluable for deepfake detection, where manipulation artifacts often manifest as subtle inconsistencies across different feature channels.

The SE mechanism operates through two fundamental operations i.e., squeeze and excitation. The squeeze operation aggregates global spatial information using global average pooling to generate channel descriptors, while the excitation operation employs a bottleneck fully connected network to model channel relationships and generate adaptive scaling factors. These scaling factors emphasize informative channels while suppressing irrelevant ones. This property is perfectly aligned with deepfake detection needs, where manipulation artifacts may be concentrated in specific feature channels.

Recent research has demonstrated SE blocks' effectiveness across multiple deepfake detection paradigms. Dasgupta et al. [26] showed how SE blocks can be strategically incorporated into lightweight CNN designs for dynamic channel-wise feature recalibration. Bhattacharjee et al. [27] developed CAE-Net by integrating EfficientNet with SE blocks, DeiT, and ConvNeXt models enhanced with Haar wavelet transformation and achieved robust generalization across diverse datasets. The channel attention paradigm has been extended to audio deepfake detection through Multi-Frequency Channel Attention (MFCA) mechanisms [28]. Balafrej et al and Decoret [29] incorporated SE blocks into steganalysis-inspired architectures, maintaining fine-grained spatial resolution while providing channel-wise attention.

Deng et al. [30] integrated SE modules into EfficientNet-V2's MBConv structures for video deepfake detection. Sheng et al. [31] developed ID-insensitive models employing SE-inspired channel compression techniques for cross-dataset generalization. gao et al. [32] combined SE block channel attention with Block-wise Discrete Cosine Transform for adaptive frequency-aware mulit-spatial attention. Wu et al. [33] achieved 99.88% accuracy on FF++ through their Interactive Two-Stream Network, applying SE-inspired attention across RGB and frequency modalities. Chen et al. [34] developed comprehensive spatiotemporal attention frameworks incorporating channel-wise attention with convolutional long short-term memory(convLSTM), achieving 99.00% AUC on FF++ dataset with superior cross-dataset generalization. Shi et al. [35] proposed Face Reconstruction-Based Generalized Deepfake Detection (FRG2D) combining SE-like attention with reconstruction learning, achieving 99.18% AUC on FF++ and 74.77% on cross-dataset evaluation.

2.5 General Deepfake detection approaches

Advancements in machine learning-based deepfake detection have progressed from single-feature analysis to increasingly complex methodologies. Researchers advanced multi-feature detection by integrating diverse face attributes, while others improved identification by biological signal analysis spanning spatial-temporal dimensions utilizing facial landmarks [36, 37]. A notable method investigates the combination of artifacts that arise when the created content is reintegrated into the frame, resulting in the creation of CNN models that use edge detection techniques [38]. A potential technique examines the correlation between facial traits and their contextual environment, as evidenced by frameworks that compare foreground and background information while allowing networks to autonomously identify distinguishing features

[39]. Moreover, GAN-generated deepfakes often exhibit unique fingerprints that can be recognized through examination of these subtle patterns and characteristics. Despite these machine learning algorithms attaining accuracy levels of up to 98%, their detection accuracies is not consistent as it depends on the characteristics of the dataset and the selection of features. Cross dataset testing reportedly drops accuracy to around 50%, highlighting the difficulty of creating generalizable detection methods [6].

Deep learning has driven recent advances in Deepfake detection, with Convolutional Neural Networks (CNNs) being the predominant models. Early approaches focused on detecting GAN-induced artifacts and extracting features from RGB data[40].The development of inception module-based networks (Meso-4 and MesoInception-4) proved to be a key development for video-based deepfake detection [16]. Studies indicate that deep CNN's surpass shallow architectures in supervised learning contexts.

Common deep learning models such as XceptionNet, GoogleNet, VGG, ResNet, EfficientNet, MobileNet, are used for deepfake detection purposes [41]. Feature extraction has been a pivotal area of interest, with academics developing techniques for the analysis of handcrafted features, spatiotemporal patterns, and face landmarks. Innovation persisted with the advent of capsule networks and ensemble learning methodologies, attaining accuracy rates beyond 99%. Recurrent Neural Network (RNN) models have been also used for deepfake identification. RNN-based methodologies have enhanced detection of deepfakes with the use of feature extraction at both micro and macroscopic scales. Researchers have devised optical flow approaches and autoencoder designs to mitigate overfitting issues [42].

With the enhancement of deepfake quality, novel architectures incorporating attention modules, such as Vision Transformers (ViT), were developed [43]. The identification of deepfakes mostly relies on the minute characteristics of facial photographs. The implementation of attention techniques for fake face identification has yielded advantages. Recent advancements have greatly enhanced detection skills via multimodal analysis, that integrates both audio and visual information [16]. The use of attention-based architectures has significantly transformed this field by facilitating more sophisticated and context-sensitive detection algorithms [5].

Researchers have developed advanced detection methods by integrating CNN and ViT architectures. EfficientNet B0 has been combined with ViT to develop models for deepfake detection, which utilizes the advantages of both methodologies. Vision Transformers (ViTs) often demonstrate low performance on datasets of intermediate size when compared to CNNs and require datasets of huge size to train from to achieve optimal performance [10]. Convolutional Vision Transformers optimize the advantages of both convolutional neural networks and Vision Transformers [44].

Despite the outstanding achievements of these approaches, the area still faces issues regarding model generalization and its adaptability across varied datasets. The transition from traditional machine learning techniques to advanced deep learning methods, by incorporating attention mechanisms clearly shows the progression in deepfake detection performance, yet the pursuit of more resilient and universally applicable solutions persists.

To address this generalization issue, we proposed an ensemble models with and without an SE block using domain-enhanced training. Unlike recent studies, where [22] used two different versions of the baseline model, EfficientB0 with a variation in the activation function by replacing the regular Swish activation with ReLU in the Conv2D layers, another study[45] employed a majority voting approach combining three different models with principal component analysis (PCA) for dimensionality reduction. In that approach, majority voting is applied after obtaining the individual model performances. In contrast, our study combines two different baseline models, whose features are fused before passing them to the classifier. Specifically, we employed VGG19 due to its proven excellent feature extraction capabilities, although it is prone to the vanishing gradient problem. We combined VGG19 with ResNet50, thus the vanishing gradient issue is eliminated through its skip connections, and the model is capable of capturing deeper semantic features. Further, to improve the channel-wise attention mechanism, we integrated our fusion model with an SE block mechanism, which makes our model architecture unique compared to existing models.

3 Methodology

This section presents the comprehensive methodology employed in developing our ensemble-based deepfake detection framework. Section 3.1 describes the data collection process, including the creation of our custom CCD-FusedFaces dataset using Face Fusion tool for both FaceSwap and Expression restorer use cases and the utilization of the 140k Real-and-Fake Faces dataset for comparative evaluation. Section 3.2 details the data augmentation techniques applied to balance the dataset by employing ten different transformation methods to expand our fake image samples from 20,000 to 60,000. Section 3.3 outlines the preprocessing pipeline applied to both CCD-FusedFaces and 140k datasets, including image resizing, normalization, and various transformations required to prepare inputs for the pretrained CNN models. Section 3.4 presents the ensemble model architecture, explaining the integration of VGG19 and ResNet50 with and without Squeeze-and-Excitation blocks, resulting in five distinct model variants for comparative analysis. Finally, Section 3.5 describes the experimental setup, which includes hardware specifications, hyperparameter tuning by employing ReduceLROnPlateau method, and the domain-enhanced training strategy designed to improve cross-dataset generalization performance.

3.1 Data collection

For this study, we created our own dataset named CCD-FusedFaces using the FFHQ (Flickr-Faces-HQ) dataset as input images, which contains 70,000 real images at 1024×1024 resolution. From this, we selected approximately 1,000-1500 real images and generated 20,000 fake images using the open-source AI tool called Face Fusion*(formerly known as a next-generation face swapper and enhancer built on advanced deep learning algorithms)*, by applying two popular image manipulation techniques: *Face Swap* and *Expression Restorer*, out of possible manipulations such as Face Swap, Face Morphing, Expression Restorer, Age Modifier, head pose modification, style transfer etc.

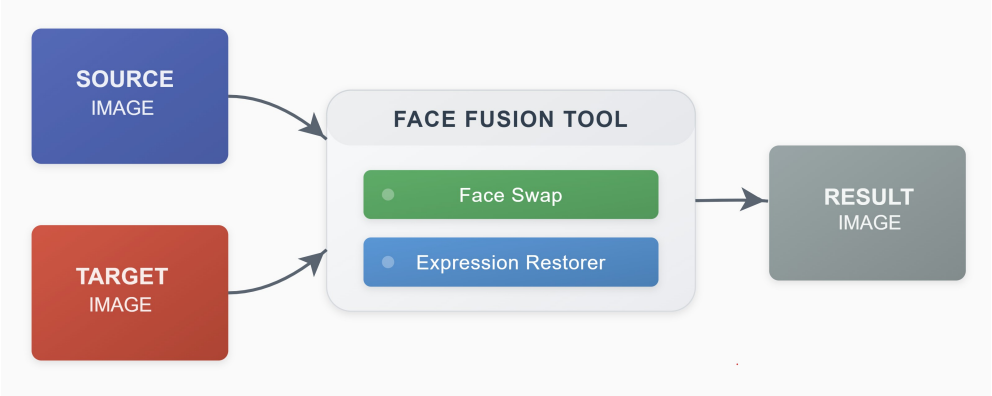


Fig. 1: Framework for image generation using face fusion tool

The FaceFusion tool we employed performs image manipulation using techniques such as face detection with 68-point facial landmarks, pretrained face swap and morph processors, Live Portrait integration, face masking, face morphing, GFPGAN (Generative Facial Prior GAN) for restoration, and ESRGAN (Enhanced Super-Resolution GAN) for super-resolution. For each use case, Face Swap and Expression Restorer generate approximately 3,500 images per demographic group (Men, Women, and Children), resulting in 10,000 images per use case and over 20,000 fake images in total. To create a complete dataset, 20,000 fake images were paired with 60,000 real FFHQ images across all demographic groups, and augmentation techniques were applied to balance the classes, resulting in 60,000 fake images for training. Thus, the CCD-FusedFaces dataset comprises the following three categories: Men and Women (including both older and younger adults), and Children (ages 0–19, including infants, school-aged children, and adolescents), representing diverse ethnic backgrounds. The images in these categories also exhibit a wide range of lighting conditions, including indoor and outdoor environments, variations in daylight, and diverse background scenes. Detailed steps for data generation, augmentation, and preprocessing are provided in Sections III-A, III-B, and III-C. Figure 1 gives an overview on the image generation process using face fusion.

3.1.1 Face Swap:

Face Swap is a popular image manipulation technique in which the face of a source person is transferred to a target person. This method replaces the target's facial identity, including features such as facial shape, skin tone, and texture-with that of the source, resulting in an image that visually appears as if the source person is present in the target scene [46]. In this manipulation method, the source person's expression is transferred to the target, meaning that if the source person is smiling, then the swapped face on the target will also smile.

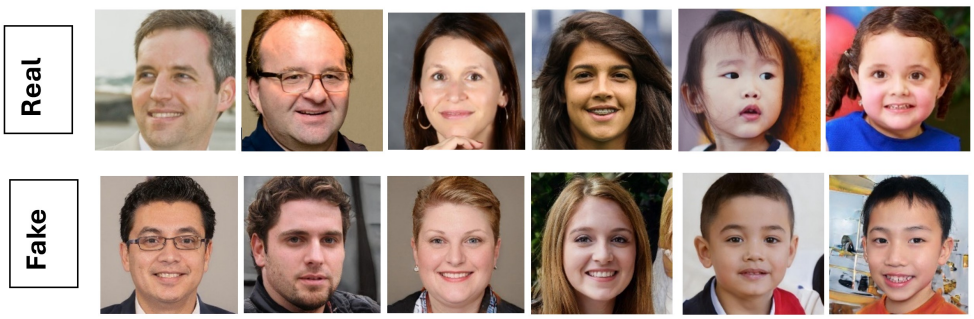


Fig. 2: Sample images from 140k-Real-And-Fake-Faces dataset

3.1.2 Expression Restorer:

Expression Restorer is a more advanced face manipulation technique, that combines face swapping with expression preservation[4]. In this case, the source face is still transferred onto the target image, just like in traditional face swap. However, instead of adopting the expression from the source, the resulting image retains the original expression of the target. This means that while the face now appears to belong to the source individual, it conveys the emotion or facial expression that was originally present in the target images such as smiling, frowning, surprise, or neutrality. Each use case listed above produced 10,000 images, resulting in 20,000 fake images across different demographic categories such as men, women, and others. To ensure consistency and realism, we processed each demographic separately, as swapping faces between vastly different features (e.g., women with bangs and men, or swapping faces of adults and children) often produced unnatural results, which were excluded. The final images were enhanced and resized to 256×256 pixels. This image generation is achieved by applying batch processing using the Face Fusion app, split into mini batches to accommodate hardware and GPU constraints. In addition to this, we utilized a publicly available dataset called 140k Real-and-Fake Faces, which consists of 70,000 real and 70,000 fake images at a resoltuion of 256x256 pixels, which is already split into training (50,000), validation (10,000), and test (10,000) sets per class.Some sample real and fake pictures from 140k-Real-And-Fake Faces datasets are shown in Figure 2.

3.2 Data Augmentation

We generated 10,000 fake images for each use case namely Face Swap and Expression Restoration, to create our own CCD-FusedFaces dataset. To form a balanced CCD dataset, we randomly selected 60,000 real images from the FFHQ dataset (which contains 70,000 images). Since we had only 20,000 fake images, we employed 10 different augmentation techniques, such as horizontal flipping and color jittering with 10 distinct scale values. Out of these ten possible augmentations, any two were applied to each fake image with an augmentation factor of 2, resulting in 60,000 fake samples. The final CCD dataset consists of 60,000 real and 60,000 fake images. This dataset

was then split into training, validation, and test sets, with 50,000, 5,000, and 5,000 images respectively for each class.

3.3 Data preprocessing

The two datasets 140k-Real-And-Fake Faces and CCD-FusedFaces are applied with following data processing techniques. The steps include torch vision transforms to prepare image inputs for the training, validation, and testing phases of the proposed ensemble model. These transformations make sure all images have the same size and format, and their pixel values are normalized. This helps the models VGG-19 or ResNet-50 to process them correctly and improves the model's performance during training. For both train and validation transforms, firstly each image is resized to 256×256 pixels to standardize its dimensions. Then, it is randomly resized and cropped to 224×224 resolution, introducing variability in object positioning and scale. This augmentation technique encourages the model to be invariant to minor spatial shifts and zooms. Then RandomHorizontalFlip and ColorJitter are also applied to add additional augmentations for further robustness, such as mirroring images and slightly altering brightness, contrast, and saturation. After these spatial augmentations, the image is converted into a PyTorch tensor, which also scales the pixel values from [0, 255] to [0.0, 1.0]. Finally, the tensor is normalized using ImageNet Mean and standard deviation values for each color channel, to ensure the data distribution aligns with the pretraining conditions of popular models like VGG-19 and ResNet-50.

3.4 Ensemble Model Architecture

Ensemble models are developed by combining multiple neural networks to achieve better performance than individual models by exploiting their diverse learning capabilities . In deepfake detection, different architectures capture distinct manipulation artifacts - while VGG19 excels at detecting low-level texture inconsistencies through its shallow feature maps, ResNet50 identifies high-level semantic anomalies using its deeper representations. This complementary behavior motivated our feature fusion approach, where both networks process the same input image simultaneously and their extracted features are concatenated before classification. The fusion strategy addresses a critical limitation of a single-model approaches: deepfakes often contain both subtle pixel-level artifacts and semantic inconsistencies that require different architectural depths to detect effectively. Our implementation creates a 2560-channel feature vector by combining 512-channel outputs from VGG19 and 2048-channel representation from ResNet50, by preserving both fine-grained details and abstract patterns. We developed five variants to systematically evaluate architectural choices: individual VGG19 and ResNet50 as baselines, basic feature fusion, fusion enhanced with SE blocks for adaptive channel weighting, and fusion with partial fine-tuning to balance pretrained knowledge with task-specific adaptation. This design enables the model to detect diverse manipulation techniques while maintaining computational efficiency through selective fine-tuning and frozen backbone weights.

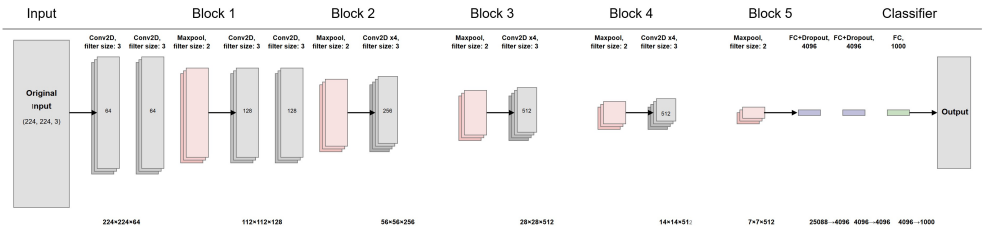


Fig. 3: VGG19 architecture

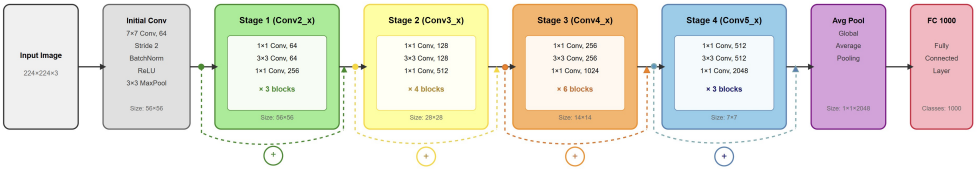


Fig. 4: Resnet50 architecture

3.4.1 VGG-19 Architecture

VGG19 is a deep convolutional neural network consisting of 19 layers i.e., 16 convolutional and 3 fully connected layers, that takes input images of size 224×224×3. It uses small 3×3 convolution filters with ReLU activations and applies max-pooling after every few layers to progressively reduce spatial dimensions[7]. The network has five convolutional blocks with increasing filter depths (64 to 512). In this study, the classification head (the fully connected layers and softmax output) is removed, and VGG19 is used solely for feature extraction. As a result, the output is a 7×7×512 tensor (after the last max-pooling layer), representing 512 high-level feature maps that capture rich semantic information from the input image. A diagrammatic representation of the VGG-19 architecture is given in Figure 3.

3.4.2 ResNet-50 Architecture

The architecture of ResNet50 model proves effective for deepfake detection due to its hierarchical feature extraction capabilities and residual learning mechanism [47]. The network's 50 layers, organized into four residual stages with bottleneck blocks, which progressively captures features ranging from low-level textures and edges to high-level semantic patterns. The skip connections inherent in ResNet50 preserves fine-grained details, that might otherwise be lost in deeper layers-a critical advantage for detecting subtle artifacts and inconsistencies introduced during face manipulation. When employed as a feature extractor in our framework, ResNet50 outputs a $7 \times 7 \times 2048$ feature tensor that encodes both the global facial structure and local texture anomalies characteristic of deepfakes. The bottleneck architecture ($1 \times 1 \rightarrow 3 \times 3 \rightarrow 1 \times 1$ convolutions) efficiently processes these features while maintaining computational efficiency

and enabling the network to identify complex manipulation patterns such as unnatural skin textures, boundary artifacts around the swapped face region, and inconsistent lighting conditions. These rich, multi-scale representations from ResNet50, when combined with complementary features from VGG19, provide our dual-stream architecture with comprehensive visual information necessary for robust deepfake detection, particularly in challenging scenarios where manipulations are sophisticated and visually convincing. A diagrammatic representation of the ResNet-50 architecture is given in Figure 4.

3.4.3 SE Block

The Squeeze-and-Excitation (SE) block serves as a crucial attention mechanism in our architecture, enabling the network to adaptively recalibrate channel-wise feature responses [25]. This module addresses a fundamental limitation in conventional convolutional operations, which treat all channels equally without considering their varying importance for the specific task at hand.

The SE block operates through a two-stage process on input feature maps of dimension $H \times W \times C$. First, the squeeze operation employs global average pooling to compress spatial information into a channel descriptor of size $1 \times 1 \times C$. This aggregation captures the global receptive field for each channel, effectively encoding the distribution of channel-wise feature responses. Subsequently, the excitation operation models channel interdependencies through a bottleneck formed by two fully connected layers. The first layer reduces dimensionality from C to C/r (where r denotes the reduction ratio, typically set to 16), followed by ReLU activation. The second layer then projects back to the original C dimensions. A sigmoid activation function produces the final channel attention weights, constraining outputs between 0 and 1.

These learned attention weights are applied to the original feature maps through channel-wise multiplication, allowing the network to emphasize informative features while suppressing less relevant ones. The skip connection ensures that the original feature information remains accessible, with the SE block learning to modulate rather than replace the input features. In the context of deepfake detection, this selective channel emphasis proves particularly valuable. Different channels may capture various aspects of facial manipulation, some responding to texture inconsistencies, others to color artifacts, and yet others to structural anomalies. The SE block enables our model to dynamically prioritize channels that are most discriminative for detecting specific types of forgeries. For instance, channels sensitive to skin texture patterns might be upweighted when detecting face-swap artifacts, while those capturing color distribution could be emphasized for detecting GAN-based manipulations. This adaptive feature recalibration, when applied to the fused features from VGG19 and ResNet50, enhances the model's ability to focus on subtle but critical indicators of facial manipulation, thereby improving detection accuracy across diverse deepfake generation techniques. The mathematical descriptions for all the operations in the SE block module are taken from [25] are given below:

The squeeze operation is as follows, in symbols:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

where, z_c is the squeezed value for channel c, H and W are the height and width of the input feature map, respectively, and $u_c(i, j)$ represents the value at position (i, j) in channel c.

The reduce operation equation is as follows, in symbols:

$$\mathbf{w} = \text{ReLU}(W_1 z + b_1) \tag{2}$$

where, $\mathbf{w}$ is the reduced dimensionality vector, W_1 and b_1 are the weights and biases for the reduction layer, z is the input channel descriptor from the squeeze operation, $ReLU$ refers to the rectified linear unit activation function.

The excitation equation is as follows, in symbols:

$$s = \sigma(W_2 w + b_2) \tag{3}$$

where, s is the channel-wise recalibration weight vector, W_2 and b_2 are the weights and biases for the excitation layer, σ denotes the sigmoid function.

Lastly, the scaling operation is as follows, in symbols:

$$v_c(i, j) = s_c \cdot u_c(i, j) \tag{4}$$

where, $v_c(i, j)$ is the value at position *(i,j)* in channel c of the output feature map, and s_c is the recalibration weight for channel c. A simple visual flowchart and a detailed diagrammatic representation of the SE block attention architecture are given in Figure 5, 6, respectively.

3.4.4 Ensemble Model for Feature Extraction and Classification:

In this study, we performed an ablation analysis with five model variants to assess how architectural changes affect binary classification performance in deepfake detection. The five models include two baseline CNNs namely; Model-1 (VGG19) and Model-2 (ResNet50) and three fusion-based ensemble models :

- Fusion Model (VGG19-ResNet50)
- Fusion Model with Attention (VGG19-ResNet50-SE Block)
- Fusion Model with Partial Fine Tuning (VGG19-ResNet50-Partial Fine-Tuned)

We employed PyTorch framework for model implementation. The two ensemble models were designed by combining the complementary feature representations of VGG19 and ResNet50. In the Fusion Model with SE Block, an attention mechanism (Squeeze-and-Excitation block) is introduced after feature fusion to recalibrate channel-wise feature responses. In the Fusion Model with Partial Fine-Tuning, the ResNet50 is partially fine-tuned by unfreezing its last 20 layers, and thus allowing it to adapt more

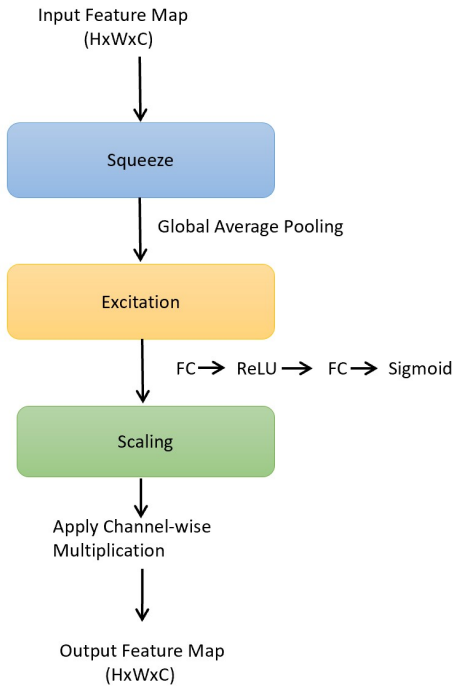


Fig. 5: SE Block architecture flowchart representation

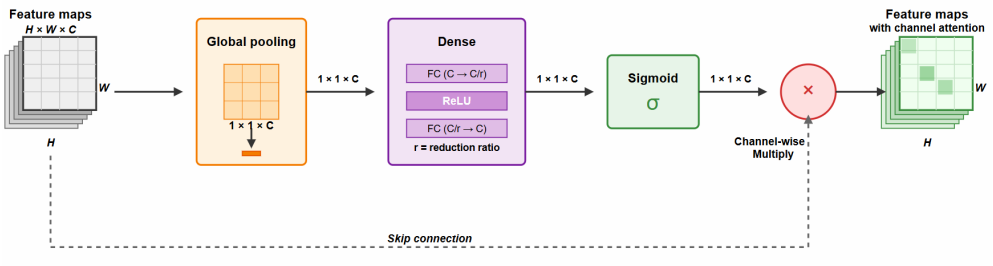


Fig. 6: SE Block architecture

effectively to the target domain while retaining the general features learned during pretraining. For both baseline models, VGG19 and ResNet50 are used solely as feature extractors. Their pretrained weights are frozen, and extracted features are passed to a custom classifier for final prediction. In the proposed fusion model, feature maps from VGG19 and ResNet50 are concatenated along the channel dimension, forming a 2560-channel representation (512 from VGG19 and 2048 from ResNet50). For a batch of input images, VGG19 produces feature maps of shape (batch_size, 512, 7, 7), and ResNet50 produces (batch_size, 2048, 7, 7) after its final convolutional layers. To enable concatenation, the spatial dimensions are aligned, with VGG19's feature

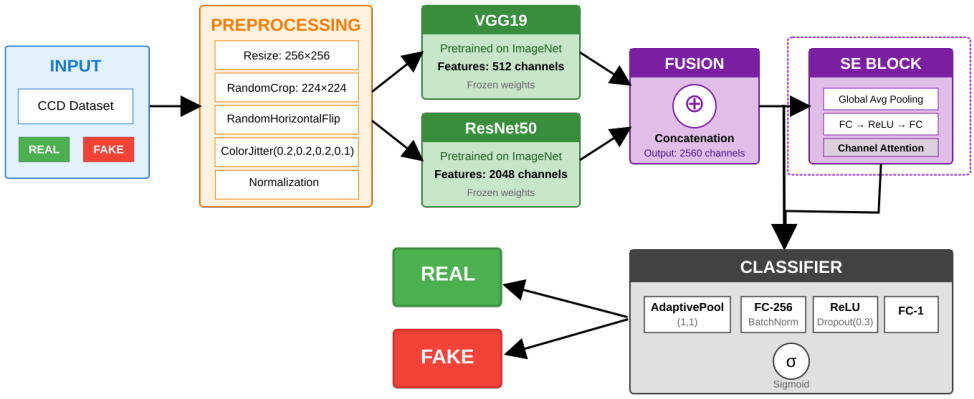


Fig. 7: Overall system architecture

maps resized via bilinear interpolation if needed to match ResNet50. The concatenation along the channel axis yields a fused feature map of shape (batch_size, 2560, 7, 7), preserving spatial structure while integrating complementary features from both networks. The fused representation is subsequently pooled and fed into the classifier for prediction. For the attention-based variant, the fused feature map is passed through an SE block to enhance channel-wise dependencies before classification. In contrast, the basic fusion model (without attention) feeds the concatenated features directly into the downstream layers. All fusion models use adaptive average pooling to compress spatial dimensions to 1×1, followed by a flattening operation to prepare the features for classification. A custom classifier is implemented as a Multilayer Perceptron (MLP) using a feed-forward architecture, shared across all models, and is built using PyTorch's Sequential module. It includes a Flatten layer, a Linear transformation from reduces the dimensionality from the number of input features (self.fusion_channels) to 256 units, Batch Normalization to stabilize training, a ReLU activation for non-linearity, a Dropout layer (30%) to reduce overfitting, another Linear layer mapping to a single output, and a final Sigmoid activation to produce a probability score for binary classification. This overall design leverages the fine spatial detail captured by VGG19 and the deep semantic understanding of ResNet50. Through careful architectural variation, such as the addition of attention and selective fine-tuning, the study aims to optimize feature representation for robust deepfake image detection. Figure 7 gives an overview of the system architecture developed for detection of fake images.

3.5 Experimental Setup

The experiment is conducted using NVIDIA's RTX A5000 GPU configured with four GPUs, equipped with 256 GB of system RAM. Each GPU has 24 GB of GPU RAM. This section further discusses the hyperparameters used for the experiments and the setup for model training and evaluation.

3.5.1 Hyper Parameter Tuning:

We applied manual hyperparameter tuning in this binary classification task. BCEWithLogitsLoss function is employed, which combines a sigmoid activation with binary cross-entropy loss. BCEWithLogitsLoss is preferred when compared to a separate Sigmoid layer followed by BCELoss, as it avoids potential instability during training. The optimizer chosen is Adam with a learning rate of 0.0001, applied only to the model parameters that require gradient updates, this is particularly important when using pre-trained models where some layers are frozen. To further refine learning, a learning rate scheduler, ReduceLROnPlateau is employed. This scheduler lowers the learning rate by 50%, by monitoring the validation loss, if the loss does not improve for three consecutive epochs and helps the model converge more effectively by adapting the learning rate during plateaus in training. The maximum number of epochs are set to 200 with an Early Stopping criteria and patience=10, which monitors validation loss and halt training when no improvement is seen for 10 consecutive epochs, helping to prevent overfitting. Table 1 provides a summary of the hyperparameters used in the experiment.

Table 1: Hyperparameters used for the model training

Hyperparameter	**Value**	**Description**
Loss Function	BCEWithLogitsLoss	Combines sigmoid activation with binary cross-entropy loss
Optimizer	Adam	Applied only to parameters requiring gradient updates
Learning Rate	0.0001	Initial learning rate for the optimizer
LR Scheduler	ReduceLROnPlateau	Monitors validation loss for adaptive learning rate adjustment
Scheduler Factor	0.5	Learning rate reduction factor when validation loss plateaus
Scheduler Patience	3 epochs	Number of epochs with no improvement before reducing learning rate
Maximum Epochs	200	Maximum number of training epochs
Early Stopping	Enabled	Monitors validation loss to prevent overfitting
Early Stopping Patience	10 epochs	Number of epochs with no improvement before stopping training

3.5.2 Model Training and Evaluation:

The train, validation, and test sets are loaded using a PyTorch Data Loader. For each epoch, the baseline and ensemble models are trained using the training dataset and validated against the validation set until the early stopping criterion is triggered. The training and validation accuracies and losses are recorded for each epoch. Once the early stopping condition is met, the best-performing model is saved and can later be loaded from disk to test on new data. This best model is then evaluated on the test dataset, which remains unseen during training, and the test accuracy, confusion

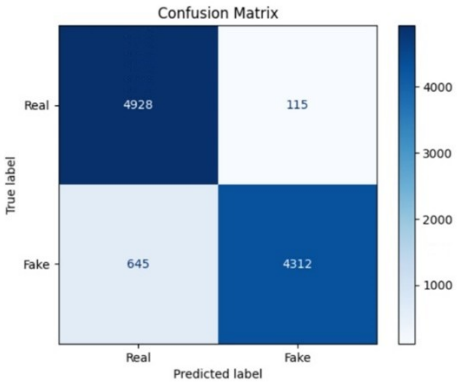


Fig. 8: Confusion Matrix for VGG-19, trained and tested on CCD-FusedFaces dataset

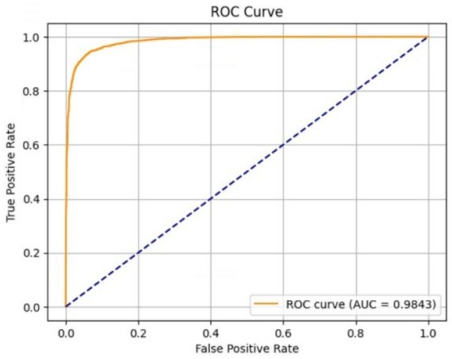


Fig. 9: ROC Curve (AUC = 0.9843) for VGG-19, trained and tested on CCD-FusedFaces dataset

matrix, and classification report, Area under the curve-Receiver Operating characteristics (AUC-ROC) are generated. To evaluate model generalization, we first loaded the model trained on the CCD-FusedFaces dataset and tested it against the test set of the 140k-Real-and-Fake Faces dataset. Similarly, we loaded the model trained on the 140k-Real-and-Fake Faces dataset and tested it against the test set of the CCD-FusedFaces. This cross-dataset validation was performed for all five models, including baseline and ensemble models. However, we observed that the models did not perform well in this setup and indicate challenges in generalizing across datasets. To address the generalization issue, we developed a domain-enhanced training methodology. In this approach, all models were trained and validated using a combination of the training and validation sets from both the 140k-Real-and-Fake Faces and CCD-FusedFaces datasets. After early stopping, the domain-enhanced best model was saved and tested on the test sets of each dataset individually, as well as on the combined test set. Across all three test configurations,both individual and combined-the model achieved higher accuracy, demonstrating improved generalization. Thus, we conclude that to enhance the generalization capability of deep learning models, it is important to train them on diverse and extensive datasets, which allows the models to perform more effectively on unseen or new data.

4 Results and Discussions

We trained and evaluated all five models using our custom CCD-FusedFaces and 140K-Real-And-Fake-Faces and tested them with both datasets. The experimental evaluation highlights the performance and generalization capability of various deep learning models, including VGG19, ResNet50, and their fused versions, trained with and without fine-tuning and SE (Squeeze-and-Excitation) blocks. Baseline models like VGG19 show moderate accuracy when trained and tested on the same dataset. For example,VGG19 achieves 92.4% accuracy on CCD but perform poorly on cross-dataset

Table 2: Performance comparison of different models across various training and testing configurations

	Training Configuration				Performance Metrics					
Model	**Train Set**	**Val Set**	**Test Set**	**Cross Dataset**	**Train Acc**	**Train Loss**	**Val Acc**	**Val Loss**	**Test Acc**	**Cross Test Acc**
VGG19	CCD	CCD	CCD	140K	0.8844	0.5621	0.9112	0.5451	**0.9240**	0.5978
VGG19	140K	140K	140K	CCD	0.7809	0.5986	0.8344	0.5767	0.8405	0.7225
VGG19	CCD+140K	CCD+140K	CCD	140K	0.8043	0.5916	0.8256	0.5766	0.8913	**0.8192**
VGG19	CCD+140K	CCD+140K	CCD+140K	-	0.8043	0.5916	0.8256	0.5766	0.8432	-
ResNet50	CCD	CCD	CCD	140K	0.9583	0.5292	0.9443	0.5124	**0.9655**	0.5804
ResNet50	140K	140K	140K	CCD	0.8650	0.5642	0.9124	0.5314	0.8882	0.7379
ResNet50	CCD+140K	CCD+140K	CCD	140K	0.8860	0.5578	0.8679	0.5543	0.9604	**0.8698**
ResNet50	CCD+140K	CCD+140K	CCD+140K	-	0.8860	0.5578	0.8679	0.5543	0.9000	-
Fusion Model	CCD	CCD	CCD	140K	0.9616	0.5281	0.9449	0.5126	0.9564	0.5771
Fusion Model	140K	140K	140K	CCD	0.8835	0.5561	0.9290	0.5331	0.9314	0.7309
Fusion Model	CCD+140K	CCD+140K	CCD	140K	0.9063	0.5487	0.9239	0.5384	**0.9583**	**0.8908**
Fusion Model	CCD+140K	CCD+140K	CCD+140K	-	0.9063	0.5487	0.9239	0.5384	0.9118	-
Fusion + Attention	CCD	CCD	CCD	140K	0.9934	0.5100	0.9793	0.4976	**0.9950**	0.5945
Fusion + Attention	140K	140K	140K	CCD	0.8806	0.5575	0.9288	0.5382	0.9230	0.7376
Fusion + Attention	CCD+140K	CCD+140K	CCD	140K	0.9021	0.5508	0.9034	0.5546	0.9743	**0.8928**
Fusion + Attention	CCD+140K	CCD+140K	CCD+140K	-	0.9021	0.5508	0.9034	0.5546	0.9199	-
Fusion + Partial FT	CCD	CCD	CCD	140K	0.9941	0.5104	0.9850	0.4959	**0.9941**	0.5557
Fusion + Partial FT	140K	140K	140K	CCD	0.9716	0.5164	0.9799	0.4943	0.9868	0.6059
Fusion + Partial FT	CCD+140K	CCD+140K	CCD	140K	0.9802	0.5142	0.9851	0.5067	**0.9956**	**0.9825**
Fusion + Partial FT	CCD+140K	CCD+140K	CCD+140K	-	0.9802	0.5142	0.9851	0.5067	0.9868	-

Table 3: Classfication Report for Different Models Across Various Training and Testing Configurations (excluding Accuracy)

	Training Configuration		Performance Metrics			
Model	Train	Test/Cross	Prec	Rec	F1	AUC
VGG19	CCD	CCD	.93	.92	.92	.984
VGG19	CCD	140K	.61	.60	.58	.598
VGG19	140K	140K	.85	.84	.84	.927
VGG19	140K	CCD	.65	.64	.63	.723
VGG19	CCD+140K	CCD	.90	.89	.89	.975
VGG19	CCD+140K	140K	.83	.82	.82	.909
VGG19	CCD+140K	CCD+140K	.84	.84	.84	.934
ResNet50	CCD	CCD	.97	.97	.97	.998
ResNet50	CCD	140K	.63	.58	.53	.661
ResNet50	140K	140K	.89	.89	.89	.956
ResNet50	140K	CCD	.77	.74	.73	.833
ResNet50	CCD+140K	CCD	.96	.96	.96	.967
ResNet50	CCD+140K	140K	.87	.87	.87	.944
ResNet50	CCD+140K	CCD+140K	.90	.90	.90	.966
Fusion Model	CCD	CCD	.96	.96	.96	.998
Fusion Model	CCD	140K	.64	.58	.52	.669
Fusion Model	140K	140K	.93	.93	.93	.983
Fusion Model	140K	CCD	.79	.73	.72	.832
Fusion Model	CCD+140K	CCD	.96	.95	.95	.998
Fusion Model	CCD+140K	140K	.89	.89	.89	.963
Fusion Model	CCD+140K	CCD+140K	.91	.91	.91	.977
Fusion Model w/ Attention	CCD	CCD	1.00	.99	.99	1.000
Fusion Model w/ Attention	CCD	140K	.62	.59	.57	.646
Fusion Model w/ Attention	140K	140K	.93	.92	.92	.981
Fusion Model w/ Attention	140K	CCD	.78	.74	.73	.827
Fusion Model w/ Attention	CCD+140K	CCD	.97	.97	.97	.998
Fusion Model w/ Attention	CCD+140K	140K	.89	.89	.89	.960
Fusion Model w/ Attention	CCD+140K	CCD+140K	.92	.92	.92	.976
Fusion Model w/ Partial FT	CCD	CCD	.99	.99	.99	1.000
Fusion Model w/ Partial FT	CCD	140K	.61	.56	.60	.622
Fusion Model w/ Partial FT	140K	140K	.99	.99	.99	.999
Fusion Model w/ Partial FT	140K	CCD	.77	.61	.54	.735
Fusion Model w/ Partial FT	CCD+140K	CCD	1.00	1.00	1.00	1.000
Fusion Model w/ Partial FT	CCD+140K	140K	.98	.98	.98	.998
Fusion Model w/ Partial FT	CCD+140K	CCD+140K	.99	.99	.99	.999

evaluations, such as dropping to 59.78% when tested on the 140K dataset. Similarly, ResNet50 trained on CCD reached 96.55% on its own test set but dropped to 58.04% in cross-dataset testing. While baseline models such as VGG19 and ResNet50 demonstrated decent performance when evaluated on the same dataset that they were trained on, their ability to generalize on cross-datasets was limited. This degradation of the performance is primarily due to the domain shift between the datasets i.e., 140k-Real-and-Fake Faces is generated using StyleGAN, whereas CCD-FusedFaces contain

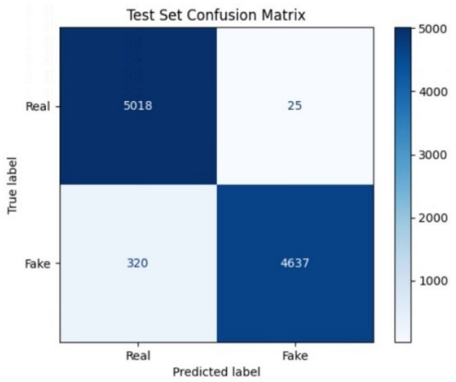


Fig. 10: Confusion Matrix for ResNet-50, trained and tested on CCD-FusedFaces dataset

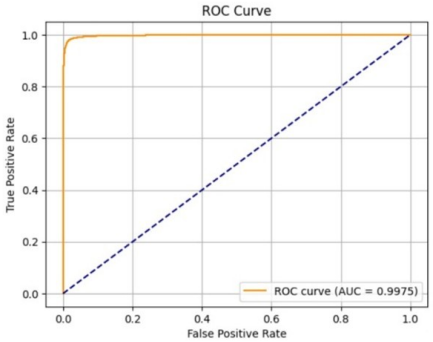


Fig. 11: ROC Curve (AUC = 0.9975) for ResNet-50, trained and tested on CCD-FusedFaces dataset

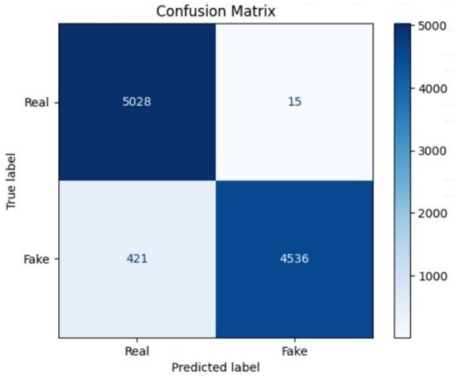


Fig. 12: Confusion Matrix for Fusion Model, trained and tested on CCD-FusedFaces dataset

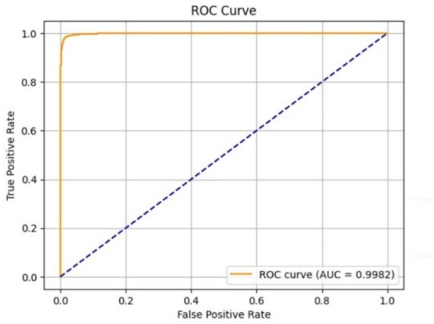


Fig. 13: ROC Curve (AUC = 0.9982) for Fusion Model, trained and tested on CCD-FusedFaces dataset

manipulations produced through a broader range of generative methods. To systematically assess generalization, we conducted a series of cross-dataset evaluations for all the models. When models trained on CCD-FusedFaces were tested on the 140k dataset, the accuracy hovered around 55-59%. Conversely, models trained with 140k achieved only around 60-73% accuracy when tested on CCD. These findings reveal that models trained on a single domain have low performance on the cross-domain dataset and they struggle to recognize fake content generated by different manipulation techniques and thus pose a significant challenge for real-world deployment. To address this, we introduced a domain-enhanced training strategy in this study. In this approach, models were trained on a combined dataset drawn from both CCD-FusedFaces and the 140k-Real-and-Fake Faces datasets. This strategy exposed the models to a more diverse

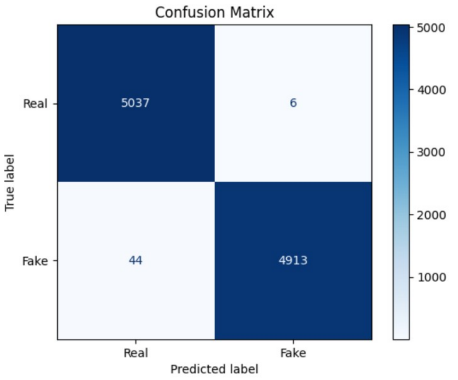


Fig. 14: Confusion Matrix for Attention-enhanced Fusion Model, trained and tested on CCD-FusedFaces dataset

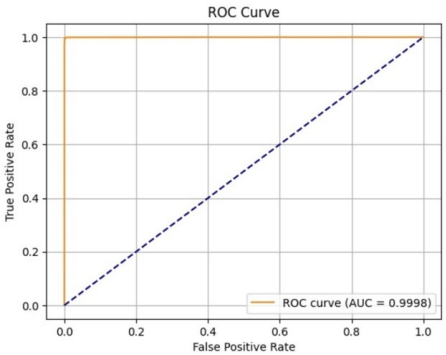


Fig. 15: ROC Curve (AUC = 0.9998) for Attention-enhanced Fusion Model, trained and tested on CCD-FusedFaces dataset

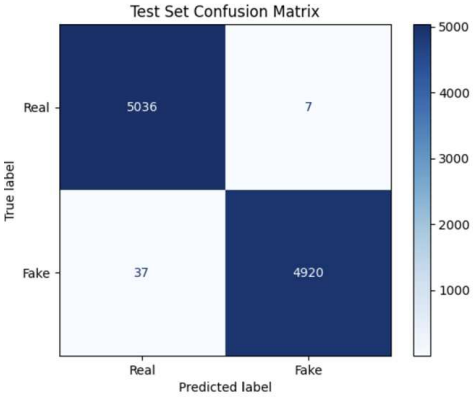


Fig. 16: Confusion Matrix for Fusion Model with partial fine-tuning, trained and tested on CCD-FusedFaces dataset

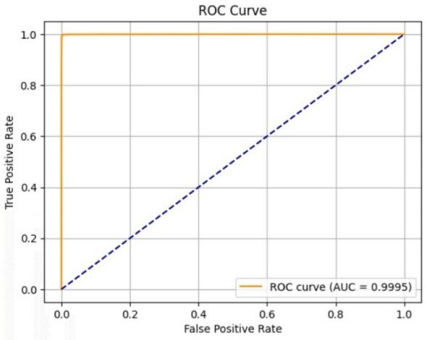


Fig. 17: ROC Curve (AUC = 0.9995) for Fusion Model with partial fine-tuning, trained and tested on CCD-FusedFaces dataset

set of fake images. As a result, models trained using domain-enhanced data achieved substantial improvements in generalization, recording training, validation, and test accuracy consistently around 81-99% for all models and specifically, Fusion model with partial fine-tuning achieved above 98%. When trained on CCD and tested on the same, ResNet50 reaches 96.55% accuracy, and cross-domain performance improves significantly when trained on a combined dataset, achieving 86.98% accuracy. The Fusion model, combining VGG19 and ResNet50, offers better results than standalone CNNs, with test accuracy up to 95.83% and improved cross-dataset accuracy up to

89.08%. Incorporating SE blocks into these fusion models further enhances performance, allowing models to recalibrate features more effectively. These Fusion Models with Attention achieve even higher scores, for example, up to 99.50% accuracy when trained and tested on CCD. When partial fine-tuning is applied to ResNet50 i.e., only the last 20 layers of the model are unfrozen, the fusion model achieves peak performance. The best-performing configuration, a fusion model with partial fine-tuning, reaches 99.41% accuracy on the CCD-FusedFaces test set and an impressive 99.56% in cross-dataset evaluation, demonstrating excellent generalization. This performance boost can be attributed to ResNet50; when it is used purely as a feature extractor, trained on the ImageNet dataset, it did not perform well on other data distributions, such as 140k-Real-And-Fake Faces and the combined dataset, due to domain shift. However, when the last few layers of Resnet50 were unfrozen, the fusion model with partial fine-tuning adapted to the current data distribution, resulting in a consistent accuracy improvement across all intra and cross-dataset testing scenarios. Moreover, the domain-enhanced training approach adopted in this study enabled the model to learn a diverse range of manipulation patterns across multiple domains, thereby achieving an accuracy of 99.56% when tested on the CCD-FusedFaces dataset, surpassing current State-of-the-Art methods.

Overall, the results clearly demonstrate that model performance and generalization are significantly improved through a combination of feature-level fusion, attention mechanisms (SE blocks), fine-tuning, and training on diverse datasets, including multiple generative techniques. These findings emphasize that such architectural and training enhancements are essential for building robust deepfake detection models that are capable of handling unseen and varied data, especially under diverse lighting conditions. A full summary of these evaluations, including individual and cross-dataset combinations, is provided in Table 2. The classification report, presented in Table 3, demonstrates strong performance on both in-domain & cross-domain evaluations.The confusion matrix for all models trained and tested on CCD is also presented below.

Based on the performance evaluation results, our proposed models demonstrate exceptional capability in distinguishing between real and AI-generated face images. The confusion matrices show consistently high classification accuracy across all architectures, with the fusion model variants achieving near-perfect performance. Especially, the Fusion Model with Partial Fine-Tuning correctly classified 5,036 real images with only 7 false negatives, while achieving 4,920 true positives for fake images with 37 false negatives. This pattern of minimal misclassification continues across all models, with Fusion Model with Attention layer showing even better performance (5,037 real correctly classified with 6 false negatives and 4,913 fake correctly classified with 44 false negatives). The ROC curves further validate the following findings, all models achieves AUC values exceeding 0.98, indicating excellent discriminative power. The fusion models with attention and partial fine- tuning mechanisms particularly stands out with an AUC of 1.0 (0.998 and 0.9995 respectively rounded off to 1.0), suggesting near-optimal separation between real and fake image classes. While VGG19 shows the lowest AUC of 0.9843, it still maintains strong performance, with its confusion matrix revealing 4,928 correctly classified real images and 4,312 correctly

classified fake images. The consistently steep initial rise in all ROC curves demonstrates that our models achieve high true positive rates while maintaining very low false positive rates, which is crucial for practical deepfake detection applications, where both precision and recall are equally important.

Additionally, the evaluation of the proposed models on the CCD dataset highlights a clear trade-off between computational efficiency and model performance as shown in Table 4 . VGG19 (80.6 MB) achieves moderate accuracy (92.40%) and reasonable inference speed (2,570 images/sec), while ResNet50 (96.5 MB) performs best on the CCD dataset with high accuracy (96.55%) and fast throughput (5,336 images/sec) due to its efficient residual architecture. Fusion-based models increase in size substantially (177–180 MB), which reduces inference speed (1,467–1,585 images/sec) but significantly improves accuracy, with Fusion with Attention and Fusion with Partial FT reaching over 99.5%. Importantly, Fusion with Partial FT generalizes well across multiple datasets, whereas ResNet50's performance drops outside the CCD-FusedFaces dataset. Efficiency analysis reveals that ResNet50 is the most efficient in terms of accuracy per megabyte(MB) and images processed per second. At the same time, fusion models prioritize high accuracy and robust generalization at the expense of speed.

These results highlight the clear trade-off between model complexity, inference speed, model performance, and generalization, suggesting that model choice should consider both real-time requirements and the need for robust detection across diverse datasets. The proposed models can effectively detect fake content designed to promote cyberbullying across demographic groups, including men, women, and children, particularly school-aged children and teenagers who frequently engage with online digital media. Future potential lies in experimenting with lightweight models, such as MobileNet, EfficientNet, XceptionNet, or ShuffleNet alongside ResNet-50, aiming to enhance inference speed while maintaining high accuracy and generalization across diverse datasets. This framework can be extended to multiclass classification of men, women, and children using our CCD-FusedFaces dataset, which we would like to address in our future experiments. Our dataset contains three demographic groups separately and is capable of achieving multiclass classification task. This multiclass classification will not only distinguish between real and fake content for each demographic group but will also help identify the most vulnerable group. This identification is necessary to create targeted rules and laws to protect victims. It can also support in enhancing the educational programs to raise awareness about the benefits and risks of digital technology and the importance of safe online practices. Furthermore, developing adversarial perturbed datasets would facilitate deeper cross-validation and robustness testing, while incorporating explainability techniques into the model would enhance transparency and interpretability, representing valuable directions for future research.

4.1 Comparison with Existing works

This section presents a comparative evaluation of recent deepfake detection models on the 140K dataset, as well as on video-based datasets such as Real and Fake Face Detection, DFDC, FaceForensics++, and CelebDF(v2), each offering a balanced mix of real and manipulated facial images and videos to enable comprehensive benchmarking. Additionally, we present our results on the CCD dataset. Since CCD is a new

Table 4: Comparison of Inference Speed, and Test Accuracy on the CCD Dataset

Model Name	**Model Size (MB)**	**Inference Speed (Images/sec)**	**Test Accuracy (CCD)**
VGG19 (Baseline)	80.6	2570	0.9240
ResNet50 (Baseline)	96.5	5336	0.9655
Fusion Model	177.1	1528	0.9583
Fusion + Attention	180.4	1467	0.9950
Fusion + Partial FT	177.1	1585	0.9956

dataset, direct comparison with prior studies using the same data is not possible. Therefore, we compare our findings with those of studies that have employed similar ensemble approaches, observing that our models outperform the state-of-the-art (SOTA) methods. The primary goal of this study is to assess model performance based on architecture type, feature extraction techniques, and classification accuracy. The results achieved in this study, highlight not only the effectiveness of individual models but also how specific design choices such as ensemble methods and deep learning techniques, contribute to improved detection accuracy. Performance metrics are summarized in Table 5 to provide a practical perspective for model selection in real-world scenarios.

Table 5: Comparative Analysis of Deepfake Detection Models

Study	**Dataset**	**Model/Arch.**	**Techniques**	**Accuracy(%)**
Alshammari & Elleithy [48]	140K	MesoNet (opt.)	Fine-tuning + weight opt.	96.20
Mallet et al. [49]	140K	MLP & LSTM	Hybrid DL model	74.7
Mallet et al. [50]	140K	CNN & SVM	Hybrid ML framework	88.33
Arshad et al. [51]	140K	Lightweight CNN	GAN edge trace	95
Dasgupta et al. [5]	140K	CNN + Multi-Head Attn.	Attention enhancement	98
Karathanasis et al. [52]	140K	Compressed CNNs	Compression + TL	Acc: ∼98
Maryam,Taeb et al. [9]	140K	CNN (6 layers)	Facial & head movement opt.	95
Debashis,Samal et al. [53]	140K	Basic CNN	Spatio-temporal features	95.18
Soudy,Ahmed Hatem, et al. [54]	FaceForensics++, DFDC	CviT	Majority voting from the ensemble model	85
Yasser,Basma, et al. [55]	FaceForensics++, DFDC	EfficientNet-B4	Deep CNN models	92.9
Yasser,Basma, et al. [55]	FaceForensics++, DFDC	XceptionNet	Deep CNN models	90.1
Xi,Aprille J., et al. [56]	Real and Fake Face Detection	SwinTransformer	Hybrid DL model+Fine tuning	71.29
Fusion model with partial FT (Ours)	**140K dataset**	**Ensemble + ResNet FT**	Feature fusion+Fine tuning	**98.68**
Fusion model with Attn.(Ours)	**CCD dataset**	**Ensemble + SE Block**	Feature Fusion+Attention	**99.50**

5 Conclusion and Future Work

This study introduces a novel image dataset, i.e., CCD-FusedFaces, comprising images generated using both face swap and expression restorer techniques, with a particular emphasis on the inclusion of diverse demographic groups, especially children. The primary objective was to develop computationally efficient deep learning ensemble models capable of detecting image deepfakes. To address this challenge, we proposed an ensemble approach that integrates channel-wise attention mechanisms (SE Blocks) with pretrained convolutional neural networks, namely VGG19 and ResNet50. The ensemble models demonstrated exceptional performance by achieving over 99% accuracy across both intra-dataset and cross-dataset evaluations and significantly outperforming many existing methods. The results underscore the models' robustness and ability to generalize across various deep-fake generation techniques. The proposed VGG19-ResNet50 ensemble approach can be adapted to audio and video deepfake detection, enabling a multimodal framework. Audio can be processed as spectrograms, and video as frame sequences with possible LSTM integration, highlighting the potential for future multimodal exploration. Additionally, this work can be further extended by training the proposed models on datasets generated using other techniques such as Autoencoders, Manual editing tools (e.g., Photoshop), Diffusion models, and additional GAN-based methods like Deep Convolutional GAN (DCGAN), Progressive GAN (PGGAN), and Wasserstein GAN (WGAN). Such domain-enhanced training would improve the applicability of these models in real-time scenarios. Broadening the range of domains represented in the training data will enhance the real-world effectiveness and robustness of deep fake detection systems, especially in safeguarding vulnerable populations within digital platforms. In real-time applications, the proposed ensemble models can be integrated into social media platforms as a pre-publication filter, automatically evaluating content and flagging or blocking identified deepfake material to help mitigate cyberbullying. Alternatively, they can be deployed as a standalone application, allowing users to input images and receive immediate feedback on whether the content is genuine or manipulated. By leveraging multiple models, these systems reduce false positives and false negatives, making detection more reliable and suitable for practical deployment, where rapid, accurate decisions are essential. Additionally, from an ethical perspective, deepfake detection plays a crucial role in mitigating online harms, such as cyberbullying and the spread of misinformation. At the same time, it is essential to address potential biases in datasets and algorithms to ensure fair performance across different demographic groups. Our findings underscore the need for transparent and accountable detection systems that support responsible deployment of AI technologies, particularly when applied in real-world scenarios affecting vulnerable populations.

6 Acknowledgement

- We sincerely acknowledge the efforts of Kushal Badal and Md Tasnim Alam for their valuable inputs and assistance with tables and diagrams, which helped strengthen the manuscript.

- Our CCD-FusedFaces dataset contains synthetic images for different demographic groups, particularly those involving minors. We acknowledge that we have adhered to the ethical principles of responsible data handling and used the images solely for conducting this research.
- We acknowledge that any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

Declarations

- Funding : This research is supported by National Science Foundation (NSF)

- Conflict of interest/Competing interests : The authors declare no conflict of interest.
- Ethics approval and consent to participate : Not applicable
- Consent for publication : The CCD-FusedFaces dataset generated in this study was developed using source images from the publicly available Flickr-Faces-HQ (FFHQ) dataset [15]. In addition, the 140k Real and Fake Faces dataset [14], which is also publicly accessible, was used for experimental purposes. As both datasets are open source and have been used in compliance with their respective terms of use, we hereby provide our consent to publish the images generated and utilized in this research.
- Data availability: CCD-FusedFaces dataset is available on request and 140k-Real-And-Fake-Faces can be accessed at https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
- Author contribution: Conceptualization, methodology, investigation, S.C., K.R.; resources, K.R., M.S.; data curation, S.C., S.L.; writing—original draft preparation, review and editing-All authors ; supervision- K.R., M.S.; funding acquisition- K.R.. All authors have read and agreed to the published version of the manuscript.

References

[1] Jasserand, C.: Deceptive deepfakes: Is the law coping with ai-altered representations of ourselves? In: 2024 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–4 (2024). IEEE

[2] Kokolaki, E., Fragopoulou, P.: Unveiling ai's threats to child protection: Regulatory efforts to criminalize ai-generated csam and emerging children's rights violations. arXiv preprint arXiv:2503.00433 (2025)

[3] Laczi, S.A., Póser, V.: Impact of deepfake technology on children: Risks and consequences. In: 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), pp. 215–220 (2024). IEEE

[4] FaceFusion Contributors: FaceFusion: Next-generation face swapping tool for images and videos. https://github.com/facefusion/facefusion. Accessed: 2025-06-17 (2024)

[5] Dasgupta, S., Badal, K., Chittam, S., Alam, M.T., Roy, K.: Attention-enhanced cnn for high-performance deepfake detection: A multi-dataset study. IEEE Access (2025)

[6] Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR) **54**(1), 1–41 (2021)

[7] Pashine, S., Mandiya, S., Gupta, P., Sheikh, R.: Deep fake detection: survey of facial manipulation detection solutions. arXiv preprint arXiv:2106.12605 (2021)

[8] Hayathunnisa, V., Kuppusamy, P., Manimaran, A.: Art of detection: Custom cnn and vgg19 for accurate real vs fake image identification. In: 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC), pp. 306–312 (2023). IEEE

[9] Taeb, M., Chi, H.: Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy **2**(1), 89–106 (2022)

[10] Sadhya, S., Qi, X.: Deepfake detection with wavelet-integrated convolutional networks. In: 2024 IEEE International Conference on Big Data (BigData), pp. 4540–4547 (2024). IEEE

[11] Yesugade, K., Jadhav, R.: Implementation of deep learning techniques for deepfake classification: A comparative study using resnet-50 and vgg16. In: 2024 IEEE Pune Section International Conference (PuneCon), pp. 1–5 (2024). IEEE

[12] Safwat, S., Mahmoud, A., Fattoh, I.E., Ali, F.: Hybrid deep learning model based on gan and resnet for detecting fake faces. IEEE Access (2024)

[13] Kumar, N., Pranav, P., Nirney, V., Geetha, V.: Deepfake image detection using cnns and transfer learning. In: 2021 International Conference on Computing, Communication and Green Engineering (CCGE), pp. 1–6 (2021). IEEE

[14] xhlulu: 140k Real and Fake Faces Dataset. Dataset containing 70,000 real faces from Flickr-Faces-HQ and 70,000 fake faces generated by StyleGAN.

[15] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410 (2019). FFHQ dataset available at: https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq

[16] Kapoor, R., Mehndiratta, V., Singh, J.: Comparative analysis of deep learning for robust deepfake detection. In: 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE), pp. 1067–1072 (2024). IEEE

[17] Yee, L.K., Hamid, I.R.A., ChaiWen, C., Abdullah, Z., Kipli, K., Foozy, C.F.M.:

Deepfake image detection using resnet50 model. In: 2024 1st International Conference on Cyber Security and Computing (CyberComp), vol. 1, pp. 80–87 (2024). IEEE

[18] Sathe, P., Sabane, V., Undale, C., Uttarkar, A., Chavhan, V., Sable, N.P., Yenkikar, A.: Deepfake image detection using yolov8. In: 2024 IEEE Pune Section International Conference (PuneCon), pp. 1–5 (2024). IEEE

[19] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I.E., Mazibuko, T.F.: An improved dense cnn architecture for deepfake image detection. IEEE Access **11**, 22081–22095 (2023)

[20] Ganguly, S., Ganguly, A., Mohiuddin, S., Malakar, S., Sarkar, R.: Vixnet: Vision transformer with xception network for deepfakes based video and image forgery detection. Expert Systems with Applications **210**, 118423 (2022)

[21] Tay, W.X., Chua, H.N., Jasser, M.B., Issa, B., Wong, R.T.: Deepdect: A facial deepfake video detection application using ensemble learning. In: 2024 IEEE 12th Conference on Systems, Process & Control (ICSPC), pp. 310–315 (2024). IEEE

[22] Minhas, S.A., Mushtaq, S., Javed, A.: Efficientnetb0 ensemble model for unified deepfakes detection. In: 2023 17th International Conference on Open Source Systems and Technologies (ICOSST), pp. 1–5 (2023). IEEE

[23] Ha, H., Kim, M., Han, S., Lee, S.: Robust deepfake detection method based on ensemble of vit and cnn. In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, pp. 1092–1095 (2023)

[24] Ilhan, I., Karaköse, M.: Mesonet-vit: Meso network and vision transformer for deepfake detection. In: 2025 29th International Conference on Information Technology (IT), pp. 1–4 (2025). IEEE

[25] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[26] Dasgupta, S., Mason, J., Yuan, X., Odeyomi, O., Roy, K.: Enhancing deepfake detection using se block attention with cnn. In: 2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), pp. 1–6 (2024). IEEE

[27] Anan, K., Bhattacharjee, A., Intesher, A., Islam, K., Fuad, A.A., Saha, U., Imtiaz, H.: Cae-net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features. arXiv preprint arXiv:2502.10682 (2025)

[28] Feng, Y.: Audios don't lie: Multi-frequency channel attention mechanism for audio

deepfake detection. arXiv preprint arXiv:2412.09467 (2024)

[29] Balafrej, I., Dahmane, M.: Enhancing practicality and efficiency of deepfake detection. Scientific Reports **14**(1), 31084 (2024)

[30] Deng, L., Suo, H., Li, D.: Deepfake video detection based on efficientnet-v2 network. Computational Intelligence and Neuroscience **2022**(1), 3441549 (2022)

[31] Sheng, Y., Zou, Z., Yu, Z., Pang, M., Ou, W., Han, W.: Id-insensitive deepfake detection model based on multi-attention mechanism. Scientific Reports **15**(1), 11168 (2025)

[32] Gao, J., Xia, Z., Marcialis, G.L., Dang, C., Dai, J., Feng, X.: Deepfake detection based on high-frequency enhancement network for highly compressed content. Expert Systems with Applications **249**, 123732 (2024)

[33] Wu, J., Zhang, B., Li, Z., Pang, G., Teng, Z., Fan, J.: Interactive two-stream network across modalities for deepfake detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(11), 6418–6430 (2023)

[34] Chen, B., Li, T., Ding, W.: Detecting deepfake videos based on spatiotemporal attention and convolutional lstm. Information Sciences **601**, 58–70 (2022)

[35] Shi, Z., Liu, W., Chen, H.: Face reconstruction-based generalized deepfake detection model with residual outlook attention. ACM Transactions on Multimedia Computing, Communications and Applications **21**(4), 1–19 (2025)

[36] Ciftci, U.A., Demir, I., Yin, L.: Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE transactions on pattern analysis and machine intelligence (2020)

[37] Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265 (2019). IEEE

[38] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010 (2020)

[39] Nirkin, Y., Wolf, L., Keller, Y., Hassner, T.: Deepfake detection based on the discrepancy between the face and its context. arXiv preprint arXiv:2008.12262 (2020)

[40] Zhang, X., Karaman, S., Chang, S.-F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2019). IEEE

[41] Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H.: Deepfake detection: A systematic

literature review. IEEE Access **10**, 25494–25513 (2022) https://doi.org/10.1109/ACCESS.2022.3154404

[42] Al-Dhabi, Y., Zhang, S.: Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn). In: 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), pp. 236–241 (2021). IEEE

[43] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[44] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)

[45] Sharma, P., Kumar, M., Sharma, H.K.: A robust ensemble model for deepfake detection of gan-generated images on social media. Discover Computing **28**(1), 41 (2025)

[46] Afroz, K., Chittam, S., Roy, K.: Understanding the threat of political deepfakes. In: 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), pp. 1–6 (2025). IEEE

[47] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[48] Alshammari, H., Elleithy, K.: Deep fake and digital forensics. In: 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0283–0288 (2023). IEEE

[49] Mallet, J., Krueger, N., Dave, R., Vanamala, M.: Hybrid deepfake detection utilizing mlp and lstm. In: 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), pp. 1–5 (2023). IEEE

[50] Mallet, J., Pryor, L., Dave, R., Vanamala, M.: Deepfake detection analyzing hybrid dataset utilizing cnn and svm. In: Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, pp. 7–11 (2023)

[51] Arshad, T., Khan, M.H., Farid, M.S.: An efficient framework to recognize deepfake faces using a light-weight cnn. In: Proceedings of the 2024 9th International Conference on Multimedia Systems and Signal Processing (ICMSSP), pp. 24–29 (2024)

[52] Karathanasis, A., Violos, J., Kompatsiaris, I.: A comparative analysis of compression and transfer learning techniques in deepfake detection models. Mathematics (2227-7390) **13**(5) (2025)

[53] Samal, D., Agrawal, P., Madaan, V., Goyal, V.: Efficient deepfake detection using convolutional neural network. In: 2025 International Conference on Intelligent Control, Computing and Communications (IC3), pp. 187–192 (2025). IEEE

[54] Soudy, A.H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., Abohany, A.A., Slim, S.O.: Deepfake detection using convolutional vision transformers and convolutional neural networks. Neural Computing and Applications **36**(31), 19759–19775 (2024)

[55] Yasser, B., Hani, J., El-Gayar, S., Amgad, O., Ahmed, N., Ebied, H.M., Amr, H., Salah, M.: Deepfake detection using efficientnet and xceptionnet. In: 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 598–603 (2023). IEEE

[56] Xi, A.J., Chen, E.: Classifying deepfakes using swin transformers. arXiv preprint arXiv:2501.15656 (2025)