Leveraging Ensemble Models for Enhanced Deepfake Detection

by

Shawna Saha

May 10, 2024

A thesis submitted to the

faculty of the Graduate School of

the University at Buffalo, The State University of New York

in partial fulfilment of the requirements for the

degree of

Master of Science

Department of Computer Science and Engineering

Acknowledgments

I would like to express my deepest gratitude to everyone who supported and guided me throughout the course of this research work.

To begin with, I am extremely thankful to my advisor Dr.Nalini Ratha for for his invaluable guidance, continuous support, and feedback. His contributions have been instrumental in completing this research thesis.

I am extremely grateful to my thesis commitee member Dr.Junsong Yuan for his insightful ideas about furthering this research.

I would also like to take this opportunity to thank my colleagues in the lab for always lending a helping hand whenever needed.

My deepest gratitude goes out to the University at Buffalo's Department of Computer Science and Engineering for supporting my research endeavors with their excellent facilities and resources.

Lastly, I would like to thank my family for their unwavering support and understanding throughout my academic journey. Their patience and encouragement have been my greatest motivation.

Table of Contents

List of Tables

List of Figures

Abstract

The proliferation of deepfake technology has raised societal apprehensions regarding the authenticity of video content circulating on social media platforms. With exponential development in the field of deep learning, the production of deepfakes has advanced to the point where distinguishing them from genuine photos or videos has become exceptionally challenging. Therefore, the need for reliable detection mechanisms to combat the ramifications of fake media is now more important than ever. In this study, we explore the evolution of deepfake media creation, and the current methodologies employed for their detection. We observe that conventional deepfake detection methods predominantly rely on a single model architecture for classification. While single models for deepfake detection offer utility, they are constrained by inherent limitations. They may lack diversity, diminishing their ability to detect nuanced manipulations and reducing accuracy and robustness against sophisticated deepfakes. Moreover, these models often struggle to generalize across diverse datasets and manipulation types, potentially leading to limited performance. Additionally, single models are prone to overfitting, becoming overly specialized to training data and compromising detection of constantly evolving deepfake variations. In response to these challenges, our research delves into the development and assessment of an ensemble of our novel EffiSwinT model aimed at refining the classification of deepfake videos. EffiSwinT seamlessly integrates a convolutional neural network with an attention mechanism, elevating the model's capacity to accurately capture spatial features and thereby enhancing its classification efficacy. Through extensive experimentation and comparative assessment, we present findings showcasing the superiority of employing an ensemble of EffiSwinT models over single models for deepfake detection, particularly in terms of precision. Furthermore, we demonstrate the critical impact of large amounts of training data covering a wide range of manipulation in improving the performance of ensemble detectors. This study adds to the continual endeav-

ors aimed at curbing the dissemination of synthetic media and preserving the integrity of visual content within the digital sphere.

Chapter 1

Introduction

The progress in artificial intelligence has facilitated the development of extremely persuasive fraudulent videos, audio recordings, and images, causing a significant overlap between what is genuine and what is artificially created. This type of deceptive manipulation of synthetic media, also known as deepfakes, presents substantial threats to individuals, businesses, and society as a whole. They diminish confidence, distort information, and have the potential to result in serious repercussions.

The word "deepfake" is derived from the combination of "deep learning" and "fake," referring to material generated using artificial neural networks. The phenomenon of deepfakes first appeared on the internet towards the end of 2017, introduced by a Reddit user known as "deepfakes" [1]. This user employed deep learning techniques to overlay the faces of famous individuals over pornographic videos. The occurrence of this event resulted in a significant amount of media coverage, which in turn caused a rapid increase in the creation and distribution of deepfake-generated material. This incident also brought attention to the potential risks associated with impersonation, identity theft, and the dissemination of false information on digital platforms. In 2018, BuzzFeed published a deepfake video showcasing former President Barack Obama [2], therefore highlighting the possible ramifications and significance of this technology.

Deepfakes can be broadly categorized into two groups depending on the targeted forged modality – audio deepfakes and visual deepfakes as shown in figure 4.2 [3]. In this research, we are focusing on the classification of visual deepfakes. Visual deepfakes are further grouped into the following types based on manipulation level:

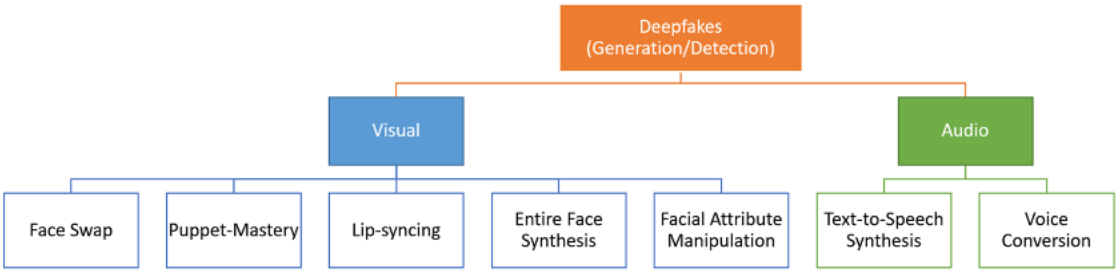


Figure 1.1: Classification of Deepfakes

- Face swap or identity swap
- Lip-syncing
- Face- reenactment or puppet-mastery
- Entire face synthesis
- Facial attribute manipulation

Audio deepfakes are further classified as:

- Text-to-speech synthesis
- Voice conversion.

There are broadly two methods for creating realistic deepfakes: generative adversarial networks (GANs) [4, 5, 6] and variational autoencoders (VAEs) [7]. GANs use two networks: the discriminator, which detects real videos, and the generator, which alters videos to trick the discriminator. VAE-based techniques use two encoder-decoder pairs. Each partner learns to dismantle and reconstruct one of the two faces involved in the exchange.

Visual Manipulation Techniques typically found in deepfake datasets are as below:

1. **Reenactment**

   - **Expression Reenactment:** Source controls the expression of a target.
   - **Mouth Reenactment (Lip-Syncing):** Target's mouth is controlled by the source or an audio input that contains speech.
   - **Gaze Reenactment:** Source determines the direction of the target's gaze as well as the position of the eyelids.
   - **Pose Reenactment:** Source controls the target's head position.
   - **Body Reenactment (Pose Transfer and Human Pose Synthesis):** Source determines the target's body posture.

     Examples of Reenactment manipulations are shown in Figure 1.2 referenced from [8].

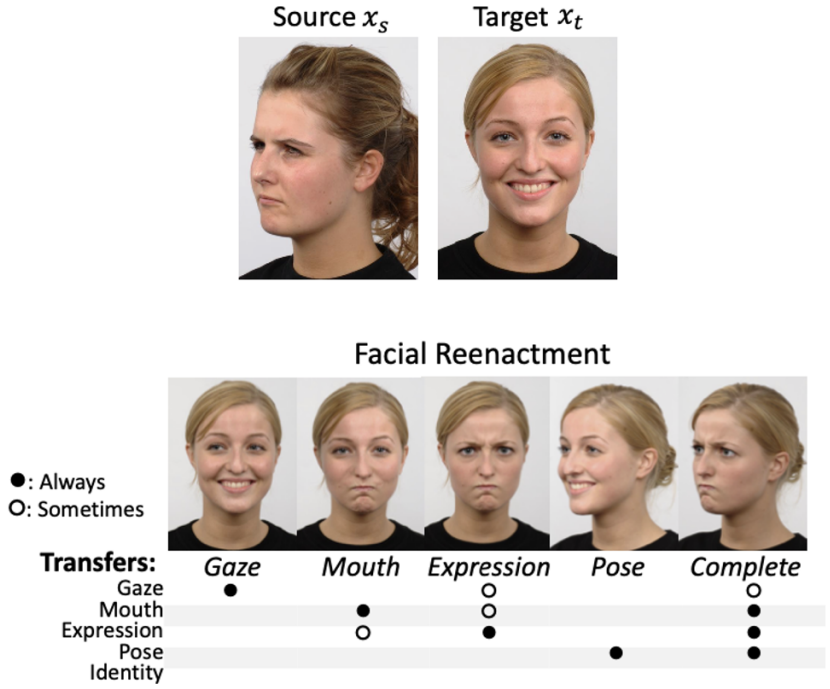


Figure 1.2: Example of Various Types of Facial Reenactment

2. **Replacement**

   - **Transfer:** Content from the source is substituted with that of the target.
   - **Swap:** Content transferred to the target from the source is driven by the target.

   Examples of Replacement manipulations are shown in Figure 1.3 referenced from [8].

Figure 1.3: Example of Various Types of Facial Replacement

3. **Editing & Synthesis**

   - **Enchantment Deepfake:** Attributes of a target are added, altered, or removed.
   - **Synthesis:** Deepfake is created with no target as a basis.

   Examples of Editing and Synthesis manipulation techniques are shown in Figure 1.4 referenced from [8].

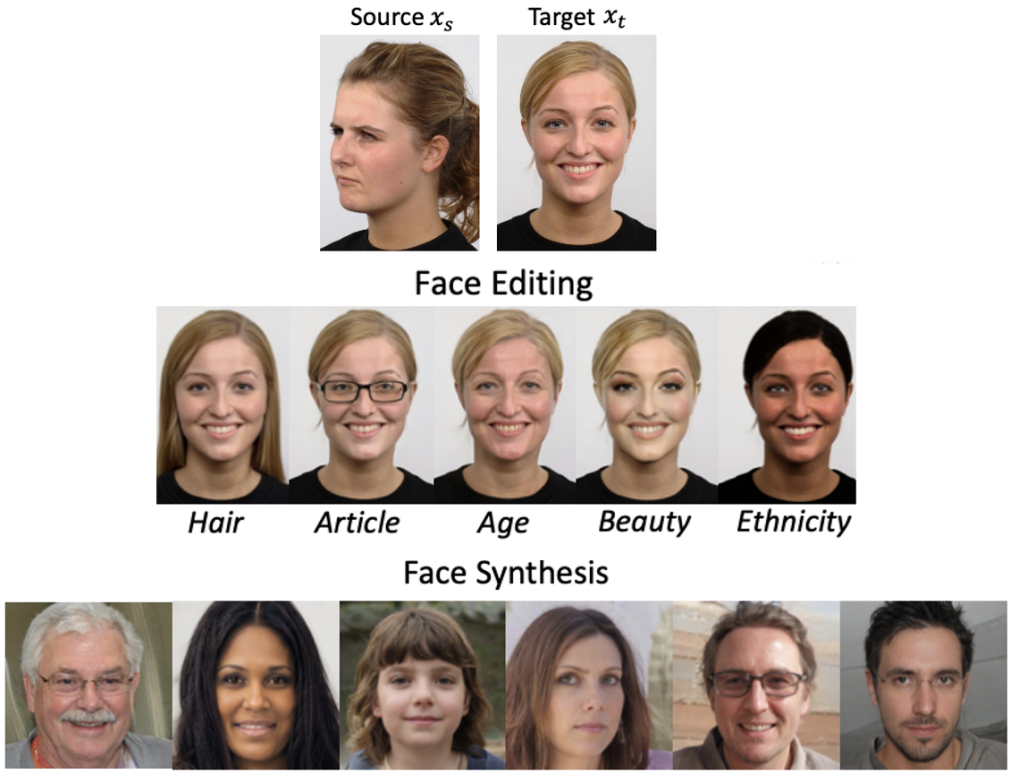


Figure 1.4: Example of Editing and Synthesis

The identification of deepfake videos may be categorized into two sections: approaches that analyze visual artifacts inside individual video frames, and methods that examine temporal aspects over several frames. While the majority of temporal feature-based approaches rely on deep learning recurrent classification models, the methods that use visual artifacts inside video frames may be implemented using either deep or shallow classifiers. Our research focuses on visual artifacts within single video frame-based methods. This technique breaks down videos into individual frames and examines visual artifacts within single frames to obtain discriminant features. These features are then distributed into either a deep or shallow classifier to differentiate between fake and authentic videos. Thus, we categorize approaches depending on the kinds of classifiers they use, namely deep or shallow classifiers.

Deep Classifiers: Deepfake videos are often created with low resolutions, which requires the use of an affine face warping approach. This technique involves scaling, rotating, and shearing the face to align it with the original configuration. Due to the difference in resolution

between the distorted facial area and the surrounding context, CNN models such as VGG16, ResNet50, ResNet101, and ResNet152 could detect artifacts created by this process. A proposed approach in [9] utilizes deep learning to identify deepfakes by analyzing the artifacts seen during the face warping phase of the deepfake generation algorithms.

Shallow Classifiers: Deepfake detection methods often rely on artifacts or inconsistencies between fake and real images or videos. Yang *et al.* [10] proposed a method observing differences between 3D head poses based on 68 facial landmarks of the central face region. The extracted features are fed into an SVM classifier for detection. Experiments on two datasets showed great performance against competing methods. Another method exploited artifacts of deepfakes and face manipulations based on visual features of eyes, teeth, and facial contours. The method exploited missing reflections, details, and texture features extracted from facial regions based on facial landmarks. Two classifiers, logistic regression, and a small neural network, were employed to classify deepfakes from real videos. However, the method requires images to meet certain prerequisites, such as open eyes or visual teeth.

Chapter 2

Related Work

Classification of deepfake videos is considered a binary classification problem. The detector's objective is to evaluate the authenticity of an image or video that contains a human face by classifying it as real or fake.

The majority of research work focussed on this domain uses Convolutional Neural Network (CNN) models for the purpose of detecting deepfake videos. The proposed studies utilize various strategies, such as augmentation techniques, temporal features combined with spatial information, recurrent networks, and transformer models, to detect deepfake images or videos. These strategies aim to enhance the models' ability to generalize. This section presents an overview of prior research conducted on the identification of deepfake videos.

In [11], the authors introduced two distinct convolutional neural network (CNN) models, namely Meso-4 and MesoInception-4, for deepfake media identification. Both of these CNN networks specifically targeted mesoscopic image features and had a limited number of layers. The authors evaluated their models on a deepfake detection benchmark and a custom dataset, achieving outstanding outcomes on both datasets.

A novel recurrent convolutional network aimed at detecting inconsistencies among neighboring frames in a video was introduced by the authors of [12]. They utilized DenseNet CNN in combination with a gated recurrent neural network (RNN) to capture both temporal and spatial features. The goal was to find any discrepancies between adjacent frames of a video that were next to each other. After conducting extensive evaluations on the widely recognized FaceForensics++ [32] benchmark, the authors obtained highly promising results in detecting deepfakes.

A deepfake detection benchmark, FaceForensics++ [13], was proposed by Rossler *et al.* in [14]. In addition to the benchmark, the authors suggested a basic XceptionNet-based CNN [15] deepfake detection method. The authors used their deepfake detection benchmark, FaceForensics++, to train and assess the basic XceptionNet. While the model did quite well when tested on the high-quality FaceForensics++ dataset's four subsets, it struggled when tested on the lower-quality videos.

Using capsule networks for deepfake detection was suggested by Nguyen *et al.* in [16]. When compared to competing methods that advocated for convolutional neural network (CNN) models, this one was unique in its proposal to use capsule networks. The detection approach that relies on capsule networks was tested on four separate deepfake datasets that include a diverse range of false movies and pictures. Statistically, the authors' suggested method outperformed competing deepfake detection methods.

In their groundbreaking work, the authors of [17] created a deepfake media detection model using convolutional neural networks (CNNs) and support vector machines (SVMs) with biological data, namely photoplethysmography (PPG) signals. A final classification score was obtained by fusing the predictions made by the CNN and SVM models. The CelebDF, Face Forensics, and Face Forensics++ datasets, among others, showed promising results when evaluated with this deepfake detection model.

Using 3D face decomposition characteristics, Zhu *et al.* in [18] presented a technique for detecting deepfakes. A combination of 3D identity texture and direct light characteristics, as shown by the authors, greatly enhanced detection performance and allowed the model to generalize well to unknown data when tested in a cross-dataset environment. The Xception-Net CNN architecture was also used for feature extraction in this investigation. To train their deepfake detection algorithm, they used both a cropped picture of a face and the 3D information that went along with it. Additionally, they thoroughly examined many feature fusion algorithms. First, FaceForensics++ was used to train the proposed model. Then, the FaceForensics++, the Google Deepfake Detection Dataset, and the DFDC [19] dataset were

used for evaluation. The model's generalizability was shown by the encouraging evaluation statistics obtained for all three datasets, which were compared to the previously described deepfake detection techniques.

For the purpose of deepfake media identification, Khan *et al.* [20] suggested using transformer architecture. The authors presented a new video-based model for deepfake identification that uses both normal cropped face photos and 3D facial characteristics for training. In addition, the authors demonstrated that their suggested model could learn from fresh data in small increments without severely losing its training material. The authors demonstrated that their suggested models performed very well on all of the datasets used to test them, including FaceForensics++, DFDC, and DFD, three of the most popular deepfake detection benchmarks.

[21] provides a Multi-modal Multi-scale Transformer (M2TR) model that scans images at several spatial levels using patches of varying sizes to detect local anomalies. M2TR employs a complex cross-modality information fusion block to better identify artifacts associated with forgery by leveraging frequency domain information in addition to RGB information. The authors demonstrate that their model beats SOTA Deepfake detection algorithms by respectable margins and prove that M2TR is successful via comprehensive trials.

The authors in [22], combined different types of Vision Transformers with a convolutional EfficientNet B0 as a feature extractor. This integration enabled them to achieve comparable results to recent methods that employ Vision Transformers. Their approach sets itself apart by not relying on distillation or ensemble techniques. Additionally, they introduced a straightforward inference procedure utilizing a voting scheme to effectively handle multiple faces in a video shot. Their experimental results demonstrated impressive performance, with an AUC of 0.951 and an F1 score of 88.0%, closely approaching the state-of-the-art performance in the Deepfake Detection Challenge (DFDC).

For the purpose of deepfake detection, an Interpretable Spatial-Temporal Video Transformer (ISTVT) was suggested in [23]. To understand spatial artifacts and temporal incon-

sistencies linked to forgeries, the suggested model integrates a self-subtract mechanism and a unique deconstructed spatio-temporal self-attention. With the help of the relevance propagation algorithm, ISTVT may additionally display the discriminative zones in terms of both time and space [46]. Extensive tests were carried out on large-scale datasets, demonstrating that ISTVT performed very well in deepfake detection inside and across datasets, proving the efficacy and resilience of the suggested model.

The authors of [24] investigated approaches to improve the ability of deepfake detection techniques to generalize by utilizing two advanced deepfake detection models, XceptionNet and EfficientNet. They utilized five databases, namely Google and Jigsaw, Face Forensics++, DeeperForensics, Celeb-DF (v2), and their own carefully curated dataset named DF-Mobio. To enhance generalization, they used many augmentation tactics during the training phase, one of which was a robust approach known as 'data farming' that entailed the utilization of random patches. In addition, they performed tests using two few-shot tuning techniques, namely fine-tuning either the first convolutional layer or the last layer of a pre-trained model, using 100 seconds of data from the training set of the test database. The findings of their tests exposed the difficulties linked to generalization in deepfake detection techniques, as the precision significantly diminished when models trained on one dataset were assessed on another. Nevertheless, the research revealed that using forceful augmentation during training and doing few-shot tweaking on the test database might improve the precision of deepfake detection in cross-database situations.

While much research has focused on developing architectures and feature descriptors for deepfake classification, there has been minimal examination of the impact of data augmentation and frame extraction on the efficacy of deepfake detection. The often used pre-processing step in deepfake classification is the extraction of individual frames from videos and the subsequent detection of facial regions. In contrast to the previously outlined approaches, we propose a technique that incorporates a combination of multi-level attention mechanisms into our innovative EffiSwinT ensemble model architecture to improve the accuracy of deepfake

classification. In addition, we use sophisticated data augmentation methods to preprocess the datasets, which leads to improved memory usage as a secondary outcome. By conducting thorough performance research on a wide range of datasets across multiple combinations of ensemble models, we demonstrate the unparalleled effectiveness of our technique.

Chapter 3

Dataset Overview

Our experiments utilize three well-known datasets, namely Face Forensics++ (FF++) [13], Face Forensics in the Wild (FFIW10K) [25], and Celeb-DF (v2) [26], for the purpose of training and evaluating our model.

Face Forensics++ is a comprehensive dataset tailored for deepfake research. It includes 1000 original video sequences along with their corresponding fake versions, created using various face manipulation and swapping techniques. The videos primarily feature unobstructed frontal faces, enabling the production of highly realistic forgeries through automated methods. The dataset covers five distinct types of deepfake attacks: DeepFakes, Face2Face, NeuralTextures, FaceSwap, with each category containing 1000 videos. Videos primarily feature frontal faces without occlusions and are sourced from 977 YouTube videos. Each subset employs different facial manipulation techniques, providing the diversity needed to train a robust model.

Face Forensics in the Wild dataset consists of 10,000 high-quality deepfake videos, with an average of three human faces in each frame. Within the context of Face Forensics in the Wild, each video comprises an average of three human faces, only certain faces are modified with selective facial manipulations. The manipulation procedure is fully automatic and is controlled by a domain-adversarial quality assessment network. This dataset allows us to evaluate the effectiveness of our algorithm on multi-person videos.

Celeb-DF (v2) dataset consists of 590 genuine videos and 5,639 deepfake videos, amounting to a total of more than two million frames. The videos have an average duration of roughly 13 seconds, with a typical frame rate of 30 frames per second. The films used

in this project are obtained from publicly accessible YouTube interviews that showcase 59 celebrities, guaranteeing a wide range of representation in terms of gender, age, and race. These films showcase a diverse array of variants, encompassing alterations in facial dimensions, orientations, lighting circumstances, and backdrops. DeepFake films are generated by exchanging facial features between the 59 individuals, resulting in videos in MPEG4.0 format. In our experiments, we employed the Face Forensics++ and Face Forensics in the Wild datasets to train and test our model on their corresponding test sets for in-dataset testing. The Celeb-DF (v2) dataset is employed for performing cross dataset testing, and we did not utilize this dataset for training our models.

Table 3.1: Comparative Analysis of the Deepfake Datasets

	Real videos	Fake videos	Synthetic Methods	Faces per frame %
FF++	1000	4000	4	1
FFIW10K	10000	10000	3	3
Celeb-DF (v2)	590	5693	1	1

Chapter 4

Methodology

Our methodology unfolds through the following sequence of steps: frame extraction, data augmentation, and model development.

4.1 Frame Extraction

Frame extraction is the process of isolating individual frames from a video. This enables analysis for signs of manipulation, discrepancies, or common artifacts in deepfake videos. However, in a video, numerous elements remain consistent across consecutive frames, evaluating each consecutive frame as an unnecessarily resource-intensive and time-consuming process. Thus, to reduce the dataset size, we implemented key-frame extraction.

"Keyframes", also known as "intra-frames" or "i-frames", are frames within a video stream that serve as representative snapshots, offering a concise and accurate summary of the video's content. Key frames indicate frames that mark the beginning or end of any transition, with subsequent frames solely containing information about transitional changes. Given our concentration on manipulated visual distortions, we believe that concentrating just on keyframes is sufficient for our algorithm to detect deepfake content. To determine the most optimal key-frame extraction threshold, we conducted experiments with threshold values of 0.3, 0.5, and 0.7. Among these, the threshold value of 0.3 demonstrated the highest level of effectiveness, leading us to select it as the standard for our process. By applying this threshold, we reduced the pre-processed data size to an average of 128 frames per video.

Furthermore, we utilized the Multi-task Cascaded Convolutional Neural Network (MTCNN)

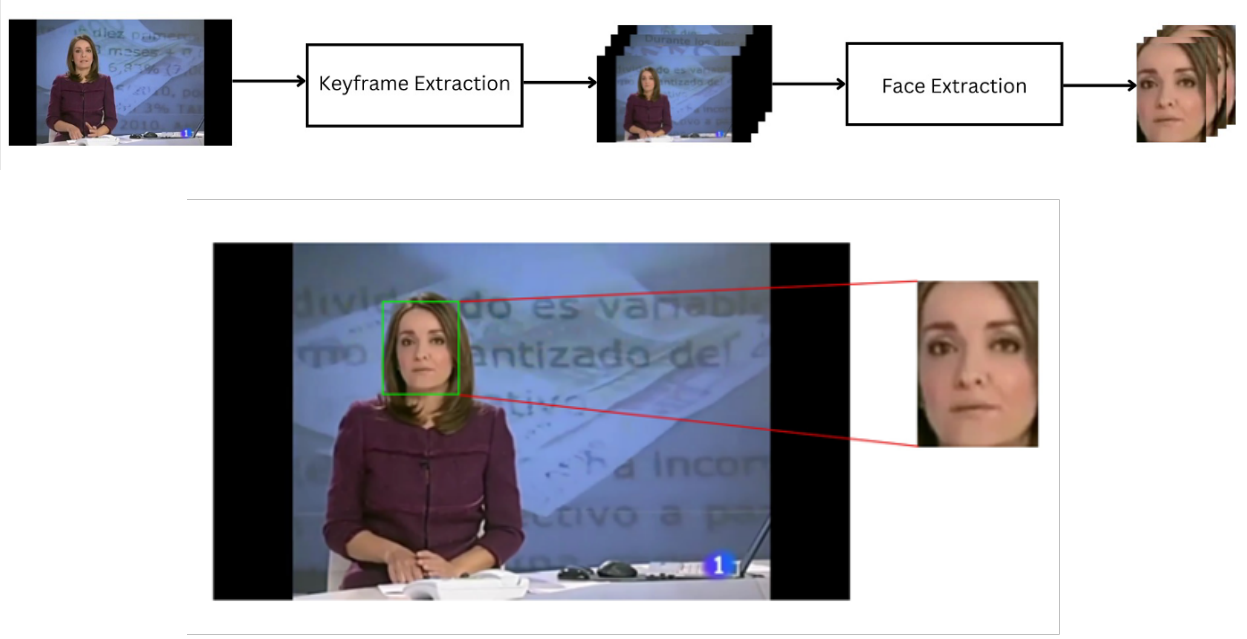


Figure 4.1: Extracting Faces from Videos

to accurately extract and store the facial regions from the key frames, ensuring high-quality data for our analysis.

4.2 Data Augmentation

A comprehensive set of data augmentation techniques was applied to the training data to enhance the robustness and generalization ability of the deepfake detection model. The data augmentation pipeline consists of a variety of transformations aimed at simulating real-world variations and distortions that may occur in facial images. These transformations applied include the following:

- **Facial Feature Dropout:** This removes facial features from the images randomly in one of the three different ways - removing the eye region, removing the mouth region, removing the nose area or keeping the image as it is. This allows for diversity in the training data.

- **Image Compression:** Image compression data augmentation technique is applied to

Figure 4.2: Example of Facial Feature Dropout Augmentation Technique

the training data with a probability of 0.2 and the compression quality is randomly chosen within the range of 60 to 100. This augmentation introduces variability in image quality by compressing images at different quality levels, which helps make the model more robust to variations in image compression that might occur in real-world scenarios.

- **Gaussian Noise:** Gaussian noise is a type of statistical noise that follows a Gaussian distribution. When applied to an image, Gaussian noise introduces random variations in pixel intensity values, mimicking the noise that can occur during image capture or transmission. This transformation was applied to each image in the training with a probability of 0.3.

- **Horizontal Flip:** This flips the image horizontally, providing additional variations of the same image.

- **Isotropic Resize:** This resizes the image while maintaining the aspect ratio. It offers three options for interpolation methods, providing variations in resizing techniques.

- **Padding:** This transformation ensures that all images, regardless of their original dimensions, are resized to have at least the specified minimum height and width, with the remaining area padded with a constant value according to the specified border

mode. This helps maintain consistency in the input size of images during training, which is crucial for deep learning models.

- **Random Brightness Contrast:** This transformation randomly adjusts the brightness and contrast of the input image. It helps in simulating variations in lighting conditions that may occur in real-world images.

- **Fancy PCA:** This transformation applies Principal Component Analysis (PCA) on the input image to perform color augmentation. It changes the color distribution in the image by altering the intensity of the principal components, thereby introducing variations in color.

- **Hue Saturation Value:** This transformation randomly adjusts the hue, saturation, and value (brightness) of the input image. It helps in simulating changes in color and intensity that may occur due to different environmental factors or camera settings.

- **To Gray:** This converts the input image to grayscale. Grayscale images contain intensity values representing different shades of gray, ranging from black to white, with no color information. Converting an image to grayscale removes color information while retaining the overall structure and luminance information. This augmentation technique introduces additional variability and diversity in the dataset, which can help improve the model's ability to generalize to unseen data and enhance its robustness against various types of input images, including those without color information.

- **Shift Scale Rotate:** This transformation introduces variations in the position, size, and orientation of the input images, which helps in making the model more robust to variations in facial expressions, poses, and viewpoints in real-world scenarios.

In the validation set only **isotropic resize** and **padding** transformations are applied.

4.3 Model Development

In this research, we propose an ensemble model, EffiSwinT, for detecting manipulated content. EffiSwinT combines EfficientNet B3 and Swin Transformer architectures.

4.3.1 EfficientNet

EfficientNet [27] is widely recognized for accurately identifying deepfake videos due to its effective compound coefficient scaling approach. Compound coefficient scaling is a unique scaling technique that equally modifies the parameters of depth, breadth, and resolution. EfficientNet demonstrates appropriate scaling by using this coefficient, making it incredibly effective in a wide range of applications, such as deepfake classification. The significant accomplishments in the Deep Fake Detection Challenge (DFDC) highlight the efficacy of pre-trained EfficientNet models, especially when enhanced by face alteration films.

Our study included examining many versions of the EfficientNet architecture, including EfficientNet B0, B3, B4, and B7. Out of all the variants, EfficientNet B3 stood out as the best performer, demonstrating better overall performance in comparison to B0. While B4 and B7 provide improved power and performance, they also need considerably more processing resources. Therefore, we chose EfficientNet B3 to achieve a favorable equilibrium between effectiveness and computing efficiency in our implementation.

4.3.2 Swin Transformer

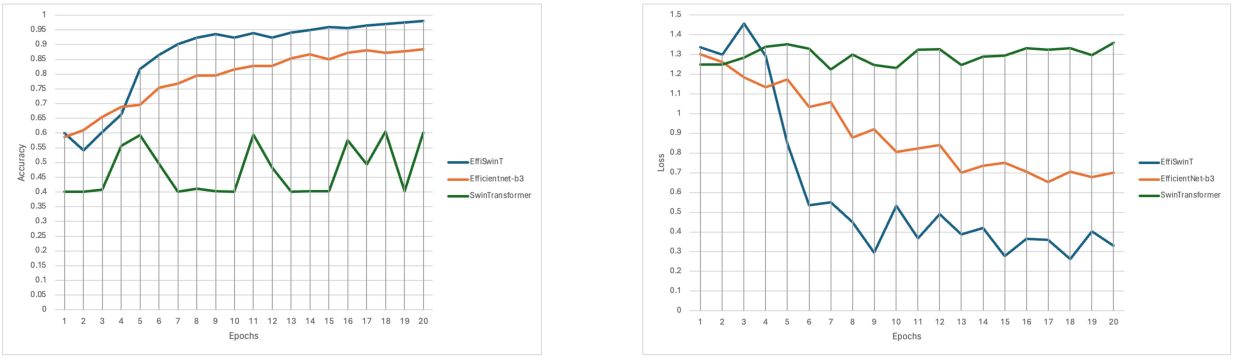
Swin Transformer [28] introduces an innovative approach to compute representations by using shifted windows. The hierarchical Transformer design improves computing efficiency by limiting self-attention to non-overlapping local windows, while nevertheless maintaining important connections between these windows. The Swin Transformer utilizes a novel architecture that efficiently captures information at different sizes, demonstrating a linear computational cost with respect to the size of the input image. Therefore, the Swin Trans-

former is a notable solution for applications that need both rapid and thorough information processing, such as image recognition and natural language comprehension.

4.3.3 Our Model Architecture

EffiSwinT integrates the strengths of both EfficientNet and Swin Transformer architectures. In our implementation of the EffiSwinT model, we have incorporated the attention heads of the Swin Transformer to handle the feature embeddings obtained from EfficientNet B3. This novel method entails omitting the conventional patch embed layer found in the Swin Transformer architecture and instead including the EfficientNet B3 as a feature extractor. This adaption not only simplifies the structure of the model but also considerably reduces its number of parameters to 13.48 million.

Figure 4.3 depicts a graphical contrast among EfficientNet B3, Swin Transformer, and EffiSwinT concerning training validation accuracy and training validation loss when subjected to training and testing on the FaceForensics++ dataset. A comparison of EffiSwinT with the state-of-the-art models in terms of the number of parameters generated is presented in Table 4.1. The figure 4.4 showcases the architecture of the EffiSwinT model along with its training pipeline.



Validation Accuracy Validation Loss

Figure 4.3: Graphical comparison between EfficientNet B3, Swin Transformer and EffiSwinT based on training validation loss and accuracy when Trained and Tested on FaceForensics++ dataset.

Table 4.1: Comparative study of the number of parameters generated by state-of-the-art models with our EffiSwinT. EffiSwinT has the lowest parameter count in millions.

Model	**Size**	**Transformer**	**Size**
DenseNet [29]	15.3 M	DeiT [30]	86.39M
XceptionNet [15]	22.9 M	ViT Hybrid [31]	98.77 M
VGG [32]	138.4 M	ViT Base [31]	86 M
Inception-v3 [33]	23.9 M	Swin Transformer [28]	27.52 M
ResNet [34]	19.4 M	Swin Transformer Attention Head [28]	1.18 M
EfficientNet [27]	10.69 M	**EffiSwinT**	**13.48 M**

4.3.4 Ensemble Models

Ensemble Learning is a methodology whereby predictions derived from multiple machine learning models are combined to yield more precise predictions compared to an individual model. Different types of ensemble learning methods include:

1. Simple Ensemble Methods:

   - **Max Voting:** Max Voting Ensemble Method involves combining predictions from multiple models and selecting the most frequent prediction as the final output. It is a simple yet effective technique used in classification tasks to improve accuracy by leveraging diverse perspectives from various models, ultimately resulting in a robust decision.
   - **Averaging:** The Averaging Ensemble Method amalgamates predictions from multiple models by averaging their outputs. After training diverse models on the same dataset, their predictions are averaged to form the final prediction. This technique helps reduce overfitting and enhances prediction accuracy through collective wisdom from diverse models.
   - **Weighted Averaging:** Weighted Averaging Ensemble Method is a technique in Ensemble Learning where predictions from individual models are combined using

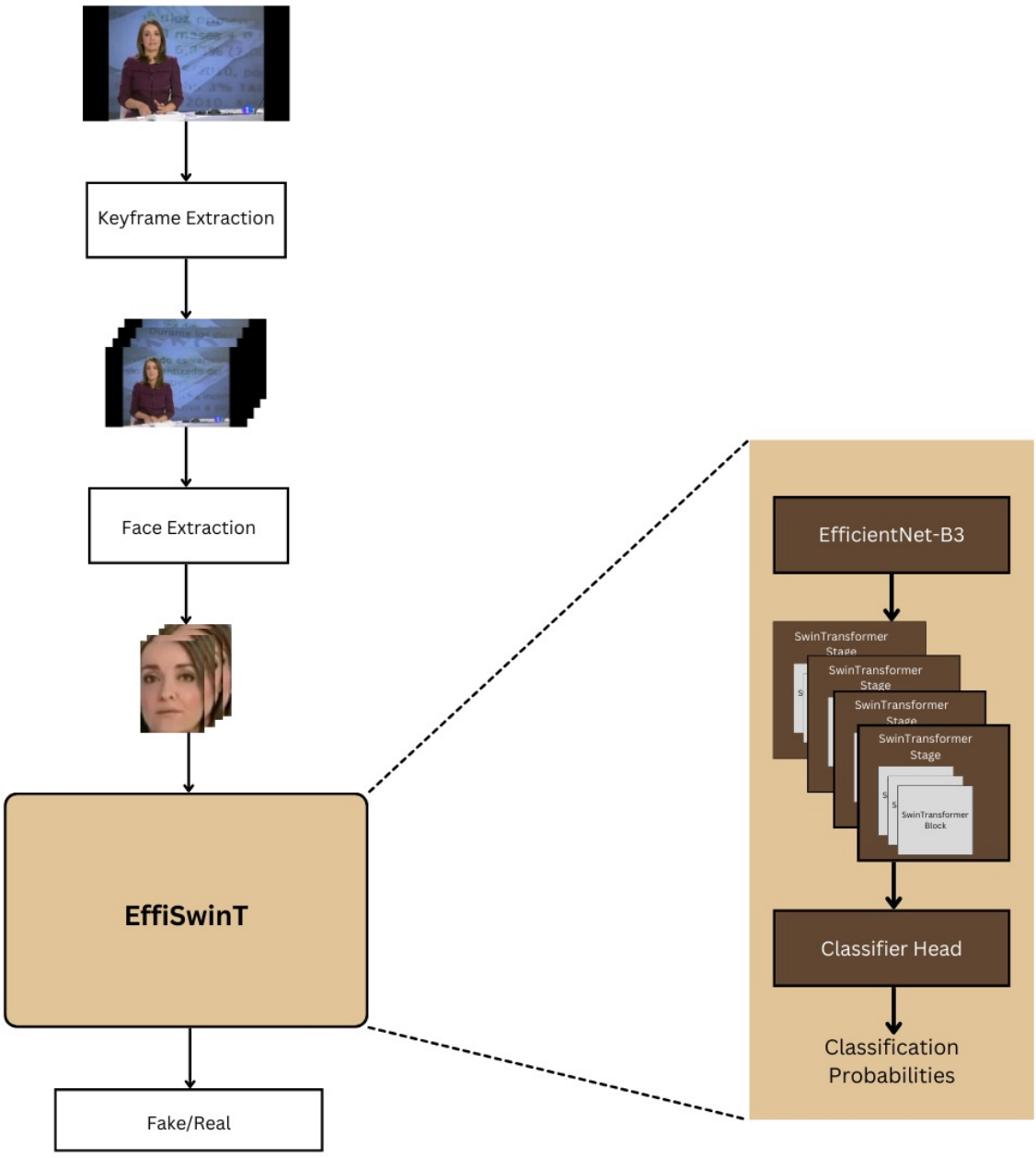


Figure 4.4: EfiSwinT Architecture

weighted averages. Each model's contribution to the final prediction is determined by its performance or expertise, enhancing accuracy by assigning higher weights to more reliable models.

2. Complex Ensemble Methods:

   - **Stacking:** Stacking Ensemble Method entails combining predictions from diverse base models using a meta-model, enhancing prediction accuracy. Initially, base models make individual predictions. Subsequently, a meta-model aggregates these predictions for the final output. This method harnesses diverse model strengths, mitigating weaknesses, and resulting in robust predictions.
   - **Blending:** Blending Ensemble Method combines predictions from diverse models by training a meta-learner on their outputs. The meta-learner learns to weigh these predictions optimally, often using a holdout dataset. This technique enhances prediction accuracy by leveraging the strengths of various models while mitigating individual weaknesses, yielding robust and accurate results.
   - **Bagging:** Bagging (Bootstrap Aggregating) is an ensemble learning technique where multiple models are trained on different subsets of the training data, sampled with replacement. Predictions are then aggregated, often through averaging or voting, to reduce variance and improve overall performance compared to individual models, enhancing robustness and accuracy.
   - **Boosting:** Boosting is an ensemble method where models are trained sequentially, with each subsequent model focusing on the errors made by its predecessors. Through this iterative process, boosting improves predictive performance by emphasizing difficult-to-classify instances, culminating in a strong learner capable of accurate predictions.

The **EffiSwinT Ensemble** architecture combines two separate EffiSwinT models, each trained on different datasets. By using a weighted average method, the predictions from

various models are combined to improve the overall accuracy of the forecasts. Every unique EffiSwinT model is trained for a duration of 50 epochs. This ensemble architecture is pivotal in harnessing the collective knowledge from diverse datasets. The model architecture of EffiSwinT Ensemble is depicted in figure 4.5.

Furthermore, our experimentation involves exploring various ensemble configurations, such as combining EffiSwinT with ResNet50, EffiSwinT with ResSwinT, and EffNetB3 with ResNet50. These variations broaden the analysis of ensemble techniques, providing insights into their effectiveness across different model combinations.

4.4 Evaluation Metrics

During our model assessment process, we examined performance in several important parameters to thoroughly analyze effectiveness. Accuracy, which is a basic metric that measures the number of true predictions out of the total number of occurrences, offers a fundamental assessment of the capability of a model. In addition, we conducted an analysis of the Area Under the Curve (AUC), which measures the model's capacity to differentiate between classes by plotting the true positive rate versus the false positive rate. These measures together allowed for a detailed evaluation of the model's performance, providing insights into both the overall accuracy of predictions and the model's ability to differentiate between different classes in classification tasks.

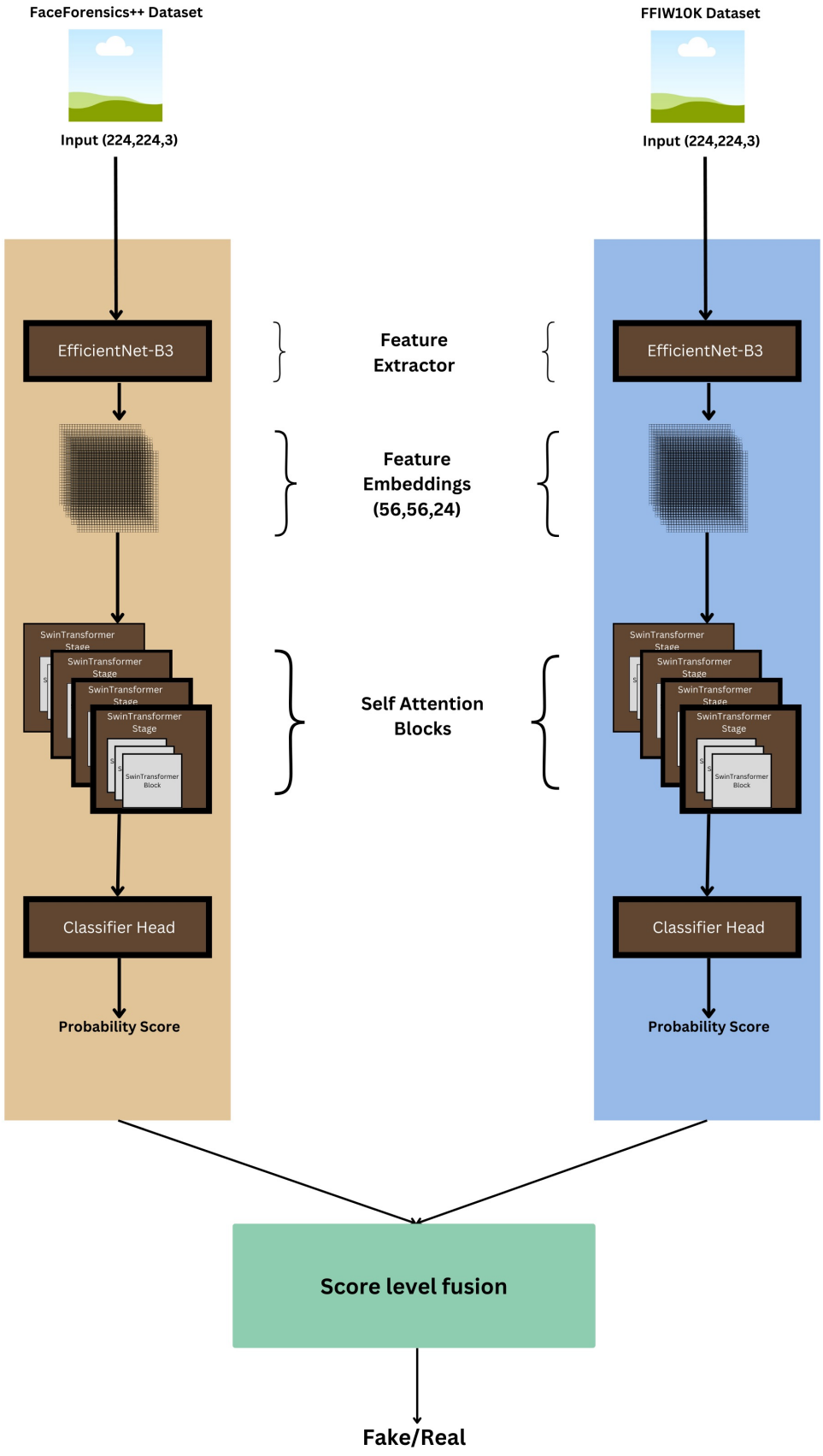


Figure 4.5: EfiSwinT Ensemble Architecture

Chapter 5

Experiments and Results

5.1 Experiments

To facilitate our analysis, we used Face Forensics++ and Face Forensics in the Wild datasets to extract essential frames from each video. Instead of using 300 frames per video, we chose keyframes from each video, resulting in an average pre-processed data size of 128 frames per video. This selection is based on a key frame extraction threshold of 0.3. During our experimentation, we tested threshold values of 0.3, 0.5, and 0.7. Among these values, 0.3 exhibited the greatest degree of optimality. Therefore, we concluded that 0.3 is the most suitable threshold value and chose to continue with it. Subsequently, each key frame is subjected to the MTCNN detector to extract and save the face area from the frames. After extracting frames from the Face Forensics++ dataset, each frame is separated into three distinct parts: training, validation, and testing. The training section consists of 720 films, while the validation and testing segments each have 140 videos. For the Face Forensics in the Wild dataset, we followed the original dataset split described in [25]. This split includes 16,000 training videos, 500 validation videos, and 3,500 test videos. The training sets are then divided into batches of 32 images and processed using a dataloader. Before inputting the batched images into EffiSwinT Ensemble models for classification, several image augmentation approaches, such as the random "Face-Dropout" augmentation, are used on the input images. Subsequently, the trained models are subjected to assessment using test sets to gauge their performance and ability to generalize. To conduct tests, we

have selected an inference threshold of 0.55 to differentiate between genuine and counterfeit films. This approach follows the same process as shown before in the figure.

In our experimental setup, we conducted a series of experiments encompassing diverse model configurations. EffiSwinT underwent training on both the Face Forensics++ and Face Forensics in the Wild datasets, followed by comprehensive evaluation on Face Forensics++, Face Forensics in the Wild, and Celeb-DF(v2). Similarly, the EffiSwinT Ensemble model underwent training on the Face Forensics++ and Face Forensics in the Wild datasets, with subsequent rigorous testing conducted on the same datasets along with Celeb-DF(v2). Subsequently, all other ensemble models were trained solely on the Face Forensics++ and Face Forensics in the Wild datasets, reserving Celeb-DF(v2) exclusively for cross-dataset testing. No models were trained on Celeb-DF(v2) dataset. This systematic approach enabled us to conduct a thorough analysis of model performance across diverse datasets, shedding light on their adaptability and efficacy in varied contexts.

5.2 Results

In our investigation, we observed that EffiSwinT demonstrated remarkable performance when trained on the FaceForensics++ dataset and subsequently tested on both FaceForensics++ and Celeb-DF(v2) datasets. However, its performance noticeably declined when tested on the Face Forensics in the Wild dataset. Conversely, when trained on the Face Forensics in the Wild dataset, EffiSwinT exhibited noteworthy performance on both Face Forensics in the Wild and Celeb-DF(v2) datasets, and its performance dropped when evaluated on Face Forensics++. The decrease in performance on the FaceForensics++ and Face Forensics in the Wild datasets may be due to the uneven nature of the FaceForensics++ dataset and the greater diversity of deepfake videos compared to Face Forensics in the Wild.

The EffiSwinT Ensemble demonstrated superior performance across all datasets in comparison to EffiSwinT. This result highlights the effectiveness of using two EffiSwinT models

trained on separate datasets and using a weighted average ensemble learning method. The ensemble technique resulted in significant improvements in accuracy and area under the curve (AUC) scores, highlighting its effectiveness in integrating diverse models to achieve improved prediction accuracy in deepfake classification. The results for both EffiSwinT and EffiSwinT Ensemble are shown in Table 5.1.

Table 5.1: Results achieved on evaluation of the EffiSwinT and EffiSwinT Ensemble model on various datasets.

Model	**Training Dataset**	**Testing**					
		FF++		**FFIW10K**		**Celeb-DF (v2)**	
		Accuracy(%)	**AUC**	**Accuracy(%)**	**AUC**	**Accuracy(%)**	**AUC**
EffiSwinT	FF++	90.23	0.958	66.86	0.846	70.46	0.730
	FFIW10K	50.95	0.698	90.50	0.942	53.86	0.531
EffiSwinT Ensemble	FFIW10K & FF++	**95.12**	**0.977**	88.83	0.941	**74.13**	**0.760**

In comparison to alternative ensemble combinations, EffiSwinT Ensemble demonstrated the best performance. Table 5.2 and Table 5.3 presents a detailed depiction of these outcomes.

5.3 Comparison with State of the Art

To demonstrate the effectiveness of our suggested EffiSwinT Ensemble model, we performed a comparative study of its accuracy and AUC findings in comparison to other cutting-edge models. The findings are shown in Table 5.4. The findings indicate that our ensemble model achieved superior performance compared to all cutting-edge models on the Face Forensics++ and Face Forensics in the Wild datasets in terms of accuracy. Furthermore, in terms of the AUC score, our model demonstrated superior performance compared to all models in the Face Forensics in the Wild dataset and all models except one in the Face Forensics++ dataset.

Table 5.5 demonstrates a comparative study of cross dataset assessment, which shows that our EffiSwinT Ensemble model performed better than the state-of-the-art models in

Table 5.2: Perfomance of different Ensemble Models when Trained on FaceForensics++ Dataset

Ensemble Model	**Training Dataset**	**Test Dataset**	**Accuracy (%)**	**AUC**
EffiSwinT + ResNet50	FF++	FF++	91.07	0.963
	FF++	FFIW10K	66.46	0.848
	FF++	Celeb-DF (v2)	72.77	0.742
EffiSwinT + ResSwinT	FF++	FF++	91.07	0.944
	FF++	FFIW10K	66.47	0.557
	FF++	Celeb-DF (v2)	72.78	0.810
EffNetB3 + ResNet50	FF++	FF++	86.30	0.913
	FF++	FFIW10K	65.13	0.562
	FF++	Celeb-DF (v2)	77.03	0.836

Table 5.3: Perfomance of different Ensemble Models when Trained on Face Forensics in the Wild Dataset

Ensemble Model	**Training Dataset**	**Test Dataset**	**Accuracy (%)**	**AUC**
EffiSwinT + ResNet 50	FFIW10K	FF++	52.98	0.619
	FFIW10K	FFIW10K	90.59	0.905
	FFIW10K	Celeb-DF (v2)	55.40	0.636
EffiSwinT + ResSwinT	FFIW10K	FF++	51.90	0.608
	FFIW10K	FFIW10K	90.65	0.906
	FFIW10K	Celeb-DF (v2)	54.63	0.628
EfficientNet B3 + ResNet 50	FFIW10K	FF++	27.14	0.238
	FFIW10K	FFIW10K	90.36	0.901
	FFIW10K	Celeb-DF (v2)	41.70	0.281

the Celeb-DF (v2) dataset, specifically in terms of AUC.

Inspired by [35], we choose to run a comparison of our Effiswint model's performance, taking into account the number of parameters created. Our EffiSwinT model demonstrated a remarkable average accuracy of 90.23% when evaluated on the Face Forensics++ dataset. It is important to mention that this great level of accuracy was achieved with a very small number of parameters, amounting to just 13.48 million. Upon comparing these findings with the data provided in Table 4.1, we can assert that our EffiSwinT model is not only very efficient in terms of CPU resources, but also surpasses other established deep learning models in accurately identifying deepfakes. This further supports the notion that our technique

achieves a remarkable equilibrium between the intricacy of the model and the effectiveness of detection.

Table 5.4: Comparison of State-of-the-art models with our EffiSwinT Ensemble on the test set of the same training dataset. Our Ensemble model outperforms all other models in terms of accuracy in Face Forensics++ and Face Forensics in the Wild dataset and AUC in case of Face Forensics in the Wild.

Dataset	**Model**	**Accuracy(%)**	**AUC**
FF++	ResNet [34]	87.04	0.992
	XceptionNet [15]	88.32	0.997
	EfficientNet [27]	87.78	0.994
	HRNet [36]	88.74	**0.999**
	VIT [31]	75.73	0.922
	BEiT [37]	86.82	0.988
	Swin Transformer [28]	87.25	0.988
	CaiT [38]	85.48	0.991
	EffiSwinT Ensemble	**90.95**	0.964
FFIW10K	XceptionNet [15]	54.1	0.561
	MesoNet [11]	53.8	0.554
	PatchForensics [39]	58.9	0.616
	FWA [40]	60.2	0.631
	EffiSwinT Ensemble	**88.83**	**0.941**

Table 5.5: Comparison of State-of-the-art models with our EffiSwinT Ensemble on the cross dataset. Our Ensemble model outperforms all other models in cross dataset validation on Celeb-DF(v2).

Model	**Training Dataset**	**Test Dataset**	**AUC**
Meso4 [11]	FF++	Celeb-DF (v2)	0.609
MesoIncep [11]	FF++	Celeb-DF (v2)	0.696
XceptionNet [13]	FF++	Celeb-DF (v2)	0.736
EfficientB4 [27]	FF++	Celeb-DF (v2)	0.748
EffiSwinT Ensemble	FFIW10K & FF++	Celeb-DF (v2)	**0.760**

Chapter 6

Conclusion and Future Work

As part of this investigation, we have suggested a hybrid model called EffiSwinT, which is a combination of the convolution model EfficientNet B3 and the transformer model Swin Transformer. This model is intended to identify deep fakes in a reliable manner. To ensure that our ensemble model could handle diverse deepfake generation techniques, we trained it on Face Forensics++ and Face Forensics in the Wild datasets and then tested it on Celeb-DF (v2) for cross-dataset evaluation and on both Face Forensics++ and Face Forensics in the Wild datasets for same dataset performance evaluation. We also analyzed the "Face-Dropout" method for data augmentation and key-frame extraction for deepfake video frame extraction. In the context of the same dataset assessment, we compared our proposed model to state-of-the-art models. The results showed that our model outperformed the state-of-the-art models in Face Forensics++ and Face Forensics in the wild dataset in terms of accuracy and area under the curve (AUC). During the evaluation process on the Celeb-DF (v2) dataset, our ensemble model demonstrated higher performance in comparison to various other models that are currently considered to be state-of-the-art.

The result is that our ensemble of EffiSwinT models, when combined with the key-frame extraction and "Face-Dropout" data augmentation strategy, offers amazing detection performance while being resource-economical in comparison to other techniques that are currently in use.

From this study, we offer a hybrid model that we name EffiSwinT for reliably detecting deep fakes. The convolutional model EfficientNet B3 and the transformer model Swin Transformer are both included in this hybrid system. By constructing an ensemble that is

comprised of two EffiSwinT models and use weighted averaging, we can improve the classification performance of our model. We trained our ensemble model on Face Forensics++ and Face Forensics in the Wild datasets to ensure that it was capable of handling a wide variety of deepfake generation techniques. After that, we tested it on Celeb-DF (v2) for cross-dataset evaluation and on both Face Forensics++ and Face Forensics in the Wild datasets for evaluation of its performance on the same dataset. In addition, we investigated the "Face-Dropout" technique for the purpose of data augmentation and key-frame extraction for the purpose of deepfake video frame separation.

When compared to the state-of-the-art models in Face Forensics++ and Face Forensics in the wild, our proposed model performed better on the same dataset in terms of accuracy and area under the curve (AUC). About the Celeb-DF (v2) dataset, our ensemble model performed much better than previous models that are considered to be state-of-the-art.

When paired with key-frame extraction and the "Face-Dropout" data augmentation technique, our EffiSwinT model ensemble exceeds state-of-the-art approaches in terms of detection accuracy and resource efficiency. This is the conclusion that can be drawn from the previous sentence. As part of this investigation, we have suggested a hybrid model called EffiSwinT, which is a combination of the convolution model EfficientNet B3 and the transformer model Swin Transformer. This model is intended to identify deep fakes in a reliable manner. We trained our proposed ensemble model on datasets such as Face Forensics++ and Face Forensics in the Wild, and then we evaluated it on the same datasets as well as a cross dataset on Celeb-DF (v2) in order to assess its resilience in identifying a variety of modification methods. Furthermore, we investigated the process of key-frame extraction, which involves the extraction of frames from deepfake films. Additionally, we investigated the method of Face-Dropout, which is used for the enhancement of data. Both with and without the deployment of the "Face-Dropout" data augmentation strategy, our model showed superior performance compared to EfficientNet. In the context of the same dataset assessment, we compared our proposed model to state-of-the-art models. The results showed that our

model outperformed the state-of-the-art models in Face Forensics++ and Face Forensics in the Wild dataset in terms of accuracy and area under the curve (AUC). Our model fared better in each of the five subsets of the Face Forensics++ dataset when it came to the area under the curve (AUC). Face Forensics++ and Face Forensics in the Wild are the two independent datasets that were used to train the ensemble model that we developed to improve the generalized performance of deepfake detection. This was accomplished by combining two different EffiSwinT architectures into the ensemble model. During the evaluation process on the Celeb-DF (v2) dataset, our ensemble model demonstrated higher performance in comparison to various other models that are currently considered to be state-of-the-art.

The result is that our ensemble of EffiSwinT models, when combined with the key-frame extraction and "Face-Dropout" data augmentation strategy, offers better detection performance while being resource efficient in comparison to other techniques that are currently in use.

In future, the scope of this research can be expanded to incorporate temporal characteristics alongside spatial characteristics. Given that EffiSwinT and EffiSwinT Ensemble are primarily image classification models, their application can be extended to a broader range of tasks. By enhancing the capabilities of these models, we can adapt them to a variety of different use cases, thereby expanding their application across diverse domains.

Bibliography

[1] Sophie Maddocks. "'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes". In: *Porn Studies* 7 (June 2020), pp. 1–9. DOI: 10.1080/23268743.2020.1757499.

[2] Cristian Vaccari and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News". In: *Social Media + Society* 6 (Feb. 2020), p. 205630512090340. DOI: 10.1177/2056305120903408.

[3] Momina Masood, Marriam Nawaz, Khalid Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward". In: *Applied Intelligence* 53 (June 2022), pp. 1–53. DOI: 10.1007/s10489-022-03766-z.

[4] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. "Deep Learning for Deepfakes Creation and Detection". In: *CoRR* abs/1909.11573 (2019). arXiv: 1909.11573. URL: http://arxiv.org/abs/1909.11573.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[6] Yisroel Mirsky and Wenke Lee. "The Creation and Detection of Deepfakes: A Survey". In: *CoRR* abs/2004.11138 (2020). arXiv: 2004.11138. URL: https://arxiv.org/abs/2004.11138.

[7] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[8] Yisroel Mirsky and Wenke Lee. "The Creation and Detection of Deepfakes: A Survey". In: *CoRR* abs/2004.11138 (2020). arXiv: 2004.11138. URL: https://arxiv.org/abs/2004.11138.

[9] Yuezun Li and Siwei Lyu. "Exposing DeepFake Videos By Detecting Face Warping Artifacts". In: *CoRR* abs/1811.00656 (2018). arXiv: 1811.00656. URL: http://arxiv.org/abs/1811.00656.

[10] Xin Yang, Yuezun Li, and Siwei Lyu. "Exposing Deep Fakes Using Inconsistent Head Poses". In: *CoRR* abs/1811.00661 (2018). arXiv: 1811.00661. URL: http://arxiv.org/abs/1811.00661.

[11] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "MesoNet: a Compact Facial Video Forgery Detection Network". In: *CoRR* abs/1809.00888 (2018). arXiv: 1809.00888. URL: http://arxiv.org/abs/1809.00888.

[12] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

[13] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1–11.

[14] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "FaceForensics++: Learning to Detect Manipulated Facial Images". In: *CoRR* abs/1901.08971 (2019). arXiv: 1901.08971. URL: http://arxiv.org/abs/1901.08971.

[15] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

[16] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos". In: *CoRR* abs/1810.11215 (2018). arXiv: 1810.11215. URL: http://arxiv.org/abs/1810.11215.

[17] Umur Aybars Ciftci and Ilke Demir. "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals". In: *CoRR* abs/1901.02212 (2019). arXiv: 1901.02212. URL: http://arxiv.org/abs/1901.02212.

[18] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z. Li. "Face Forgery Detection by 3D Decomposition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2929–2939.

[19] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. "The DeepFake Detection Challenge Dataset". In: *CoRR* abs/2006.07397 (2020). arXiv: 2006.07397. URL: https://arxiv.org/abs/2006.07397.

[20] Sohail Ahmed Khan and Hang Dai. "Video Transformer for Deepfake Detection with Incremental Learning". In: *CoRR* abs/2108.05307 (2021). arXiv: 2108.05307. URL: https://arxiv.org/abs/2108.05307.

[21] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. "M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection". In: *CoRR* abs/2104.09770 (2021). arXiv: 2104.09770. URL: https://arxiv.org/abs/2104.09770.

[22] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. "Combining efficientnet and vision transformers for video deepfake detection". In: *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*. Springer. 2022, pp. 219–229.

[23] Cairong Zhao, Chutian Wang, Guosheng Hu, Haonan Chen, Chun Liu, and Jinhui Tang. "ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 1335–1348. DOI: 10.1109/TIFS.2023.3239223.

[24] Pavel Korshunov and Sébastien Marcel. "Improving generalization of deepfake detection with data farming and few-shot learning". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (2022), pp. 386–397.

[25] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. "Face Forensics in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 5778–5788.

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics*. 2020. arXiv: 1909.12962 [cs.CR].

[27] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *CoRR* abs/1905.11946 (2019). arXiv: 1905.11946. URL: http://arxiv.org/abs/1905.11946.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[29] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention". In: *CoRR* abs/2012.12877 (2020). arXiv: 2012.12877. URL: https://arxiv.org/abs/2012.12877.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words:

Transformers for Image Recognition at Scale". In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929. URL: https://arxiv.org/abs/2010.11929.

[32] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[35] Hafsa Ilyas, Ali Javed, Muteb Mohammad Aljasem, and Mustafa Alhababi. "Fused Swish-ReLU Efficient-Net Model for Deepfakes Detection". In: *2023 9th International Conference on Automation, Robotics and Applications (ICARA)*. IEEE. 2023, pp. 368–372.

[36] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. "Deep High-Resolution Representation Learning for Visual Recognition". In: *CoRR* abs/1908.07919 (2019). arXiv: 1908.07919. URL: http://arxiv.org/abs/1908.07919.

[37] Hangbo Bao, Li Dong, and Furu Wei. "BEiT: BERT Pre-Training of Image Transformers". In: *CoRR* abs/2106.08254 (2021). arXiv: 2106.08254. URL: https://arxiv.org/abs/2106.08254.

[38] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. "Going deeper with Image Transformers". In: *CoRR* abs/2103.17239 (2021). arXiv: 2103.17239. URL: https://arxiv.org/abs/2103.17239.

[39] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. "What makes fake images detectable? Understanding properties that generalize". In: *CoRR* abs/2008.10588 (2020). arXiv: 2008.10588. URL: https://arxiv.org/abs/2008.10588.

[40] Yuezun Li and Siwei Lyu. "Exposing DeepFake Videos By Detecting Face Warping Artifacts". In: *CoRR* abs/1811.00656 (2018). arXiv: 1811.00656. URL: http://arxiv.org/abs/1811.00656.