CHAPTER I
INTRODUCTION

This chapter introduces the study's background, research questions, objectives, contributions, and limitations. It details the urgency of the topic, the core problems investigated, and the targeted outcomes. Additionally, it highlights the expected benefits while establishing the scope and boundaries of the research.

1.1. Background

The development of artificial intelligence has driven the emergence of increasingly realistic synthetic content, one of which is deepfakes. Deepfakes are digital media resulting from artificial intelligence-based manipulation that can alter or generate images, audio, and videos to make them appear authentic [1]. The urgency of this issue is highlighted by Romanishyn, Malytska, and Goncharuk, who reported that the number of deepfake videos in 2023 reached 95,820, representing a 550% increase since 2019 [2]. Diel et al., through a systematic review and meta-analysis of 56 studies, also showed that human accuracy in detecting deepfakes reaches only 55.54% [3]. These conditions demonstrate that deepfakes have become a threat to information integrity, identity security, and public trust.

Deepfake detection challenges are growing increasingly complex because manipulated media has become harder to distinguish visually. Altuncu, Franqueira, and Li explained that the diversity of generation methods, datasets, and evaluation metrics means detection cannot rely solely on shallow visual artifacts [4]. Singh and Dhumane also demonstrated that cross-dataset generalization capability and real-world conditions remain persistent challenges in deepfake detection [5]. Deploying models on mobile devices also requires architectures with relatively low computational complexity [6]. Hutagalung et al. showed that a lightweight architecture based on MobileViT with CBAM achieved 96.4% accuracy and possesses characteristics supporting mobile device implementation [7]. MobileViT itself combines local feature extraction capabilities through Convolutional Neural Networks (CNN) and global relationship modeling via transformers [8]. These

characteristics form the basis for choosing MobileViT as the main backbone in this study.

While MobileViT is capable of generating good feature representations, relying on spatial information alone may not be sufficient to capture subtle manipulation artifacts. Amen and Ranam demonstrated that integrating the Fast Fourier Transform (FFT) into a lightweight deepfake detection model achieved an accuracy of 94.2% and an F1-score of 93.8% [6]. The FFT converts image representations from the spatial domain to the frequency domain, allowing manipulation patterns that are difficult to observe visually to be analyzed based on their frequency distribution [9]. Frequency features are then used as supplementary information to complement spatial features and enhance the model's ability to recognize manipulated images. This approach also enables the model to be more sensitive to frequency artifacts that often go undetected by spatial-based methods.

Combining spatial and frequency features requires a mechanism capable of emphasizing the most relevant features. Luo and Wang, through their FMSI model, demonstrated that directional interaction between spatial and frequency features can improve cross-dataset generalization performance [10]. One mechanism that can be utilized is the Convolutional Block Attention Module (CBAM), which emphasizes critical information across channel and spatial dimensions. The application of attention mechanisms in deepfake detection is also supported by studies from AlMuhaideb et al. and Lal et al., which showed improvements in model performance and robustness [11][12]. These findings justify using Modified CBAM to help the model focus its learning process on more discriminative features. Face detection and cropping are also applied during the preprocessing stage so that classification focuses on the relevant region. Face-area-based approaches have previously been used to aid in identifying manipulated regions [13]. Lightweight YOLO-based face extractors can also support the deepfake detection process efficiently [14][15]. Based on these considerations, YOLOv11n is used to detect and crop the head area containing the face prior to classification [16]. Based on the reviewed literature, the use of lightweight architectures, frequency features, and attention mechanisms has each demonstrated distinct potential in deepfake detection. However, integrating MobileViT, FFT, and Modified CBAM into a single

lightweight classification pipeline that is evaluated on data with differing characteristics remains an open area for further development.

Based on this rationale, deepfake detection research requires not only strong classification performance, but also the ability to capture subtle manipulation artifacts, generalize effectively, and support mobile device deployment. Therefore, this study proposes a video frame-based deepfake classification model using the FaceForensics++ (C23) dataset, featuring MobileViT as the backbone, FFT for frequency feature generation, Modified CBAM as the attention mechanism, and YOLOv11n for face-area preprocessing support. Model performance is evaluated using accuracy, precision, recall, and F1-score, after which the top-performing model is deployed in an Android application prototype to evaluate its implementation feasibility on mobile devices.

1.2. Research Questions

The research problems addressed in this study are formulated into the following questions:

1. How can MobileViT, FFT-based frequency representations, and Modified CBAM be integrated into a unified facial-image-based pipeline for binary deepfake classification?
2. How does the integrated MobileViT, FFT, and Modified CBAM model with the optimal configuration perform on the secondary dataset, and how well does it generalize to the primary dataset based on accuracy, precision, recall, and F1-score?
3. How can the model with the strongest overall performance be deployed as an Android application prototype for deepfake detection?

1.3. Research Objectives

In accordance with the research questions, this study pursues the following objectives:

1. To develop a unified facial-image-based deepfake classification pipeline that combines MobileViT, FFT-based frequency representations, and Modified CBAM.

2. To evaluate the performance of the integrated MobileViT, FFT, and Modified CBAM model with the optimal configuration on the secondary dataset and assess its generalization capability on the primary dataset based on accuracy, precision, recall, and F1-score.
3. To deploy the model with the strongest overall performance in an Android application prototype for deepfake detection.

1.4. Research Contributions

This study is intended to contribute to both the scientific development and practical implementation of deepfake detection. The contributions are described as follows:

1. From an architectural perspective, this study investigates a facial-image-based deepfake classification pipeline that combines MobileViT, FFT, and Modified CBAM. The integration provides an approach for jointly utilizing lightweight visual feature extraction, frequency-domain information, and attention-based feature refinement within a single model.
2. This study is expected to provide a scientific reference regarding the performance of the integrated MobileViT, FFT, and Modified CBAM model on the secondary dataset and its generalization capability on the primary dataset based on accuracy, precision, recall, and F1-score. Therefore, the findings of this study may serve as a reference for the development of deepfake detection methods that consider not only performance on the dataset used during the research process but also the model's ability to handle unseen data.
3. From a practical perspective, the study demonstrates the implementation of a lightweight deepfake detection model within a mobile application prototype. This implementation provides an example of how deepfake detection can be brought closer to end users as an accessible means of supporting digital-content authenticity verification.

1.5. Research Limitations

To keep the investigation aligned with its intended scope, the study is conducted under the following limitations:

1. The preprocessing stage exclusively utilizes YOLOv11n for face detection and cropping.
2. Deepfake detection in this study is limited to images or video frames containing a single complete face as the primary object.
3. Deepfake detection is limited to binary classification between real and fake classes based on visual information from images or video frames.
4. This study focuses on the integration of MobileViT as the feature extractor, spatial-frequency feature fusion using FFT, and Modified CBAM as the attention mechanism within a single deepfake classification pipeline.
5. The secondary classification dataset consists of FaceForensics++ C23, which includes 1 real category named Original and 5 fake categories: Deepfakes, Face2Face, FaceShifter, FaceSwap, and NeuralTextures.
6. The primary dataset utilizes real videos sourced from YouTube and deepfake-manipulated videos generated using NeoRefacer.
7. The Android implementation is developed as an application prototype intended to examine the feasibility of deploying the deepfake detection model on a mobile device.