CHAPTER I
INTRODUCTION

This chapter comprehensively establishes the research background, detailing the problem statement, objectives, and the rationale behind the necessity and viability of the proposed solution. It further specifies the core research problems and the targeted milestones. Concluding the chapter is an outline of the study's potential contributions to academia, industry, and society at large, alongside the identified research limitations.

1.1. Background

Pulmonary diseases constitute one of the leading causes of mortality worldwide. According to the World Health Organization [1], conditions such as chronic obstructive pulmonary disease (COPD), pneumonia, and asthma rank among the top ten global causes of death. The early detection of respiratory anomalies plays a pivotal role in improving patient recovery prognoses and alleviating the burden on healthcare systems. However, conventional diagnosis via manual auscultation remains heavily reliant on the expertise and clinical experience of medical practitioners, thereby potentially introducing subjectivity and diagnostic inconsistencies [2].

In Indonesia, acute respiratory infections (ARI) remain one of the leading causes of morbidity, particularly among children [3], while environmental factors, air quality, and lifestyle continue to exacerbate the incidence of these cases [4] [5]. Diagnostic methods such as chest radiography, CT scans, and bronchoscopy remain essential modalities, yet they necessitate specialized facilities and incur high costs, making them difficult to implement widely across diverse healthcare settings [6]. This situation highlights the necessity for an alternative diagnostic system that is more efficient, non-invasive, and affordable. Several studies have demonstrated that digital respiratory sound analysis utilizing deep learning-based methods can be leveraged both at the signal level to classify normal lung sounds and adventitious breath sounds, abnormal sounds not present in healthy lungs, such as crackles and wheezes [7] [8], and at the clinical level to detect respiratory pathologies or diseases from auscultation recordings [9]. Consequently, respiratory sound analysis has emerged as a highly promising approach for computer-aided diagnosis.

The advancement of Artificial Intelligence (AI) technologies presents substantial opportunities for developing Computer-Aided Diagnosis (CAD) systems based on respiratory sound signals [10] [11]. Utilizing digital stethoscopes, pulmonary audio signals can be recorded and converted into visual representations, such as spectrograms or Mel-spectrograms, which illustrate frequency distribution over time. Distinctive acoustic patterns within these representations, notably crackles and wheezes, serve as critical indicators of various pulmonary disorders and have been leveraged as the foundational basis for automated lung sound classification [12] [13]. Consequently, this approach holds the potential to render the diagnostic process significantly more objective and rapid, while simultaneously facilitating seamless integration into telemedicine frameworks.

Previous research has demonstrated that pulmonary diseases can be detected via respiratory sound analysis. Bardou et al. [13] employed Convolutional Neural Networks (CNNs) to classify normal and pathological pulmonary sounds based on Mel-spectrograms. Furthermore, Minami et al. [14] and García Ordás et al. [9] developed CNN architectures and variational autoencoders to detect respiratory pathologies across large-scale datasets, whereas Stasiakiewicz et al. [15] utilized a combination of wavelet packet decomposition and support vector machines (SVM) to identify crackles. Fraiwan et al. [16] and Yu et al. [17] concluded that these machine learning and deep learning-based approaches consistently classify diverse respiratory diseases from pulmonary audio recordings with promising accuracy and sensitivity.

The majority of these methods predominantly rely on CNN architectures or hybrid CNN-RNN models. Asatani et al. [2] proposed a spectrogram-based Convolutional Recurrent Neural Network (CRNN) evaluated on the ICBHI 2017 dataset, reporting a sensitivity of 0.63 and a specificity of 0.83. Furthermore, Wang et al. [18], [19] developed hybrid models based on ResNet and other lightweight neural networks that concurrently process Mel-spectrograms and raw waveform signals. Despite their robust performance, CNN and CRNN models primarily extract local features, presenting inherent limitations in capturing global spatio-temporal relationships within time-frequency representations.

More recent approaches leverage the Vision Transformer (ViT), which employs self-attention mechanisms to learn inter-patch relationships across the entire

log-Mel spectrogram image. Ariyanti et al. [20] introduced the Audio Spectrogram Vision Transformer (AS-ViT) evaluated on the ICBHI 2017 dataset, reporting an unweighted average recall of up to 86.4% and an overall score of approximately 69.3% in the abnormal respiratory sound detection scenario, thereby outperforming several prior CNN-based methods. Employing a different approach, Mang et al. [7] utilized cochleograms with time-frequency representation that closely approximates human auditory characteristics, and integrated them with ViT for a four-class breath sound classification. This achieved an accuracy of approximately 67.9%, consistently outperforming CNN architectures such as VGG16 and ResNet50. The findings from these studies indicate that the self-attention mechanism within ViT is highly effective in capturing global inter-patch relationships in the time-frequency domain, demonstrating the capability to surpass the performance of purely CNN-based architectures.

The efficacy of deep learning models is significantly influenced by the quality and distribution of the training data. The ICBHI 2017 Challenge dataset exhibits a pronounced class disparity in sample distribution, wherein certain classes are substantially more dominant than others [17] [21]. This condition inherently risks inducing model bias toward the majority class, thereby degrading sensitivity to the pathological classes that are of paramount clinical significance. A prevalent strategy employed to mitigate this bias is data augmentation, the process of artificially generating additional training samples to enhance feature diversity and rectify the class distribution balance [22] [23].

Within the audio domain, data augmentation can be executed through both manual and synthetic approaches. Manual augmentation applies deterministic transformations to the signal, such as time stretching and pitch shifting, utilizing limited parameter ranges. This process aims to increase sample variations in breath segments that have been clinically annotated by physicians while preserving the primary acoustic characteristics relevant to the clinical labels [2]. Although this approach is straightforward and easy to implement, it remains limited in its capacity to represent the natural diversity of respiratory sounds. Conversely, synthetic augmentation based on generative models, such as conditional Generative Adversarial Networks and variational autoencoders, is capable of significantly enhancing signal variation [23] [24]. MelGAN [25] [26] stands as a prominent

generative model specifically designed to reconstruct and synthesize audio signals from Mel spectrogram representations. Furthermore, it has been reported to successfully generate signals demonstrating stable and natural frequency characteristics [26] [27].

Several studies indicate that generative augmentation can enhance the classification performance of medical audio signals [23] [24]. However, limited research has systematically compared the efficacy of manual augmentation against synthetic augmentation based on MelGAN [26] for respiratory sound classification utilizing Vision Transformer architectures on the ICBHI 2017 dataset [7] [20]. This study focuses on the classification of four respiratory sound classes (normal, crackles, wheezes, and both) by employing two primary augmentation strategies, specifically manual augmentation and synthetic augmentation based on MelGAN, with the Vision Transformer (ViT) serving as the classifier. The objective is to mitigate the class disparity within the ICBHI 2017 dataset by increasing data variance and to evaluate the impact of each augmentation strategy on accuracy, sensitivity, and specificity. The results are anticipated to provide empirical evidence regarding the most effective augmentation strategy for improving ViT proficiency in recognizing abnormal respiratory sounds. Concurrently, this work aims to establish a foundational baseline for the development of medical audio classification systems utilizing transformer architectures for the noninvasive early detection of pulmonary anomalies.

1.2. Problem Formulation

Drawing upon the previously elaborated background, the primary research problems that constitute the focus of this study can be formulated as follows:

1. How can the Vision Transformer (ViT) method be applied to classify four respiratory sound classes within the ICBHI 2017 dataset based on log Mel spectrogram representations, specifically normal, crackles, wheezes, and both?
2. What is the impact of two audio data augmentation approaches, specifically manual augmentation and synthetic augmentation based on the MelGAN generative model, on data variance, class distribution, and the performance of the Vision Transformer (ViT) model in classifying four respiratory sound classes within the ICBHI 2017 dataset?

1.3. Research Objectives

The objective of this study is to analyze and compare the classification performance of pulmonary diseases from respiratory sounds by applying data augmentation strategies, specifically manual augmentation, synthetic augmentation based on MelGAN, and a combination of both, utilizing the Vision Transformer (ViT) model. This research utilizes the ICBHI 2017 dataset, represented in the form of log Mel spectrograms, to categorize respiratory sounds into four classes, specifically normal, crackles, wheezes, and both. Through this comparative analysis, the study anticipates identifying the most effective augmentation strategy for enhancing data variance and balance, thereby improving the proficiency of the ViT model in supporting the early detection of pulmonary diseases based on respiratory sounds.

1.4. Research Benefits

In alignment with the previously articulated research objectives, several anticipated benefits can be derived from this study. The specific benefits of this research are outlined as follows:

1. To design and implement a classification system for pulmonary diseases derived from respiratory sounds, based on the Vision Transformer (ViT) architecture, utilizing log Mel spectrogram images from the ICBHI 2017 dataset as input.
2. To provide empirical evidence regarding the impact of two audio data augmentation strategies, specifically manual augmentation and synthetic augmentation based on MelGAN, on data variance, class distribution balance, and the performance of the ViT model in the classification of four respiratory sound classes.
3. To provide quantitative comparative results and recommendations for the most effective augmentation strategy in the classification scenario of pulmonary diseases from respiratory sounds, thereby serving as a reference or benchmark for the development of similar methodologies in subsequent research.

1.5. Scope and Limitations

In alignment with the previously articulated problem formulation, several research limitations have been established to ensure that the scope remains focused and does not become overly broad, maintaining a strict direction toward the core issues to be addressed. The specific limitations of this research are outlined as follows:

1. The dataset utilized constitutes secondary data sourced from the International Conference on Biomedical and Health Informatics (ICBHI) Scientific Challenge 2017 database, comprising audio recordings of patient respiratory sounds.
2. The scope of this research is restricted to the classification of respiratory sound categories based on the labels of normal, crackles, wheezes, and both within the ICBHI 2017 dataset. Furthermore, this study does not diagnose specific pulmonary diseases, does not determine disease severity, and is not intended to substitute for the clinical interpretation of medical professionals.
3. The data augmentation methods utilized are restricted to two primary approaches, specifically manual augmentation through transformations such as pitch shifting and time stretching, alongside synthetic augmentation utilizing MelGAN to generate artificial data variance. Additionally, a combination of manual and synthetic augmentation is implemented to evaluate the classification outcomes across several distinct scenarios.
4. This research does not evaluate alternative architectures such as CNN, CRNN, or hybrid models. These models are exclusively discussed as references from prior research, without any direct experimental comparison.