CHAPTER V
CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

Based on the results of the implementation, testing, and analysis, the Vision Transformer (ViT-B/16) model is successfully applied to classify four respiratory sound patterns within the ICBHI 2017 dataset, namely normal, crackle, wheeze, and both. The audio recordings are represented in the form of log-Mel spectrogram images sized at 224 × 224 pixels, which are subsequently utilized as inputs for the model employing ImageNet pre-trained weights and partial fine-tuning on the last six encoder blocks. The primary evaluation is conducted utilizing a patient-wise split to ensure that all segments originating from the same patient are exclusively allocated to a single subset. Consequently, the testing results more accurately reflect the model's capability in classifying data from patients never encountered during the training process.

The testing results demonstrate that augmentation strategies and data partitioning ratios significantly influence the model's performance. Among all the experimental combinations encompassing original data, manual augmentation, MelGAN synthetic augmentation, and hybrid augmentation across three patient-wise split ratios the best results are obtained through the MelGAN synthetic augmentation utilizing a 70:15:15 ratio. This configuration yields an accuracy of 68.76%, a macro sensitivity of 57.90%, a macro specificity of 85.50%, and a Macro-F1 of 59.99%. These results indicate that the MelGAN synthetic augmentation provides the most optimal performance balance across the classes within the specific configuration of this research. Nevertheless, increasing the volume of data through augmentation does not invariably improve all metrics, as the outcomes remain inherently influenced by the quality of the augmented data, the distinct characteristics of each class, and the specific patient composition within each respective subset.

Based on the class-level results, the model achieves a sensitivity of 85.15% for the normal class, 46.72% for the crackle class, 54.12% for the wheeze class, and 45.61% for the both class. The normal class emerges as the most easily recognized category, whereas the crackle and both classes remain more difficult to classify. A portion of the crackle samples is still predicted as normal, while the both class is frequently confused with the normal, crackle, and wheeze classes. These findings

indicate that the model is not yet fully capable of differentiating acoustic patterns that share overlapping characteristics, particularly within the minority classes and the combined class. Consequently, the model's performance is not yet uniformly distributed across all classes, even though the optimal configuration has yielded a higher Macro-F1 metric compared to the other experimental combinations.

The supplementary analysis utilizing the 70:15:15 segment-level split yields an accuracy of 67.92%, a macro sensitivity of 57.59%, a macro specificity of 87.12%, and a Macro-F1 of 58.74%. Compared to the patient-wise split, this method solely produces a higher macro specificity, whereas the accuracy, macro sensitivity, and Macro-F1 are slightly lower. Furthermore, the same patient may appear across different subsets, even though no identical segments exist between them. Therefore, the primary results of the research continue to rely on the patient-wise split, as it provides a far more rigorous evaluation of generalization. Overall, this research provides empirical evidence regarding the impact of augmentation strategies and data partitioning on the performance of the ViT-B/16 architecture in respiratory sound classification. However, the outcomes remain limited to sound pattern classification based strictly on dataset annotations; they cannot yet be deployed as an actual clinical diagnostic system without further thorough validation utilizing external data and more diverse clinical conditions.

5.2. Recommendations

Based on the research findings and the identified limitations, it is recommended that future studies utilize alternative respiratory sound datasets as a means of external validation. Testing on data originating from distinct patients, recording devices, environments, and varying signal qualities is necessary to ascertain the model's generalization capability beyond the ICBHI 2017 dataset. Furthermore, the evaluation should ideally be conducted using patient-wise k-fold cross-validation or multiple random seed variations to ensure that the stability of the model's performance in response to shifts in patient composition can be analyzed more comprehensively.

Subsequent development should also be directed toward enhancing the quality and selection process of synthetic data. Although the MelGAN synthetic augmentation within the 70:15:15 patient-wise split yields the best results in this research, several synthetic samples still exhibit artifacts and uneven energy patterns.

Future studies could evaluate other generative models such as conditional GANs, diffusion models, or advanced audio generative methods that are more capable of preserving the distinct acoustic characteristics of respiratory sounds. Finally, the generated synthetic data should undergo a rigorous quality control process based on spectral similarity, audio quality, and label consistency prior to its incorporation into the training set.

Strategies for handling class imbalance also require further development, particularly to enhance the recognition of the crackle, wheeze, and both classes. Beyond data augmentation, subsequent research could implement focal loss, class-balanced loss, more adaptive class weighting, or class-based sampling strategies. Error analysis must also be deepened by examining the log-Mel spectrogram characteristics, signal quality, segment duration, and frequency patterns of the misclassified samples. This approach is expected to help elucidate the underlying reasons why crackle samples are frequently predicted as normal, and why the both class continues to be confused with other classes.

Future studies could directly compare the ViT-B/16 with other architectures such as CNN, CRNN, ResNet, EfficientNet, or the Audio Spectrogram Transformerutilizing identical data partitioning and evaluation protocols. The hyperparameter optimization space can also be expanded by testing the batch size, the number of unfrozen encoder blocks, dropout rates, patch sizes, and learning-rate scheduler strategies. The evaluation must continue to prioritize Macro-F1 as the primary metric, complemented by accuracy, sensitivity, specificity, precision, per-class F1-score, and the confusion matrix, ensuring that the model's performance on minority classes is not overshadowed by the majority class. If the model is directed toward actual clinical application, testing must proceed utilizing more diverse clinical data and involve validation by medical professionals, as the model's output in this research remains strictly limited to respiratory sound pattern classification and cannot yet be utilized as a definitive basis for diagnosing pulmonary diseases.