

CHAPTER V CLOSING

5.1 Conclusion

Based on all of the results of the study "Application of PSO-Random Forest in Heart Disease Classification with Boruta Feature Selection and SMOTE-Tomek Data Balancing Technique" using the UCI Heart Disease dataset containing 920 samples, 16 features, and 5 classes, the conclusions can be drawn as follows:

1. The Baseline model tends to be stable as a benchmark with an average F1 score of 0.8435 for binary classification scenario and 0.5942 for multi-class classification scenario, but the model still has difficulty classifying due to 16 noisy and minority features in class 4.
2. The application of the Smote-Tomek data balancing technique improves the F1 metric in binary and multi-class classification, as well as addresses imbalance in the minority class. This application is suitable for use in research that has imbalanced data.
3. Applying the Boruta feature selection technique can reduce noise in less important features, although it can worsen model performance when dealing with imbalanced data. This technique has proven effective when using balanced data.
4. The Smote-Tomek data balancing and Boruta (Hybrid) feature selection techniques all performed well on multi-class classification, but performed mediocre in binary classification compared to their computational time. This implementation is optimal for multi-class classification.

5.2 Suggestions

Based on all of the results of this study, there are also several suggestions for further research that can be carried out as follows:

1. This study used a comparison of three data sharing ratios (70:30, 75:25, and 80:20) using cross-validation to analyze the effect of the ratios on model performance. Therefore, in future research, it is recommended to apply k-fold or repeated k-fold to determine their impact on the stability of the parameters generated by PSO.

2. The model used in this study focuses on Random Forest optimized by PSO with Smote-Tomek data balancing technique and Boruta feature selection technique, so that in future research it is recommended to explore further models such as the application of XGBoost or LightGBM models with Bayesian Optimization-based hyperparameter optimization either as a replacement or comparison of PSO. This suggested approach has the potential to accelerate the convergence process and improve the F1 score, especially in multi-class classification.
3. The data balancing technique used was only Smote-Tomek, so future research is recommended to implement more class-specific strategies, such as applying the Adaptive Smote technique only to the smallest minority class, or comparing it with other techniques such as Smote-Enn or ADASYN. These suggested techniques reduce the risk of overfitting due to excessive synthetic samples in the majority class.
4. The dataset that being used in this study is still limited to the UCI Heart Disease dataset, which has 920 samples with 16 features and 5 imbalanced classes. Therefore, in future research, it is recommended to conduct validation tests between datasets such as Cleveland, Hungary, Switzerland, and VA Long Beach. Furthermore, it is also recommended to add or merge datasets with the highest severity samples, so that handling data imbalance does not only rely on oversampling techniques but also through expanding the sample of the actual population data.