Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Enhancing Heart Disease Prediction through SMOTE-ENN Balancing and RFECV Feature Selection

Sabrina Putri Aulia[1]*, Basuki Rahmat[2], Achmad Junaidi[3]

[1,2,3]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur
21081010048@student.upnjatim.ac.id[1], basukirahmat.if@upnjatim.ac.id[2], achmadjunaidi.if@upnjatim.ac.id[3]*

Abstract

Heart disease is the leading cause of mortality worldwide, exerting a significant influence on the national economic burden and productivity. The identification of heart disease is imperative for the prevention of more severe conditions, as it facilitates the detection of risks and symptoms at an early stage. The development of disease prediction models using machine learning has been extensively researched; however, the field continues to encounter challenges, including uneven data distribution and the presence of large, complex datasets. The proposed solution to these issues is the optimization of the Random Forest algorithm through the integration of the Synthetic Minority Over-sampling Technique and Edited Nearest Neighbor (SMOTE-ENN) with Recursive Feature Elimination and Cross-Validation (RFECV). The objective of these methods is to address the issue of data imbalance and to reduce irrelevant features, thereby enhancing the performance of the prediction model. The combination of SMOTE-ENN and RFECV consistently produces higher recall up to 0.984 and an optimal F1 score of 0.938. These results suggest that combining SMOTE-ENN data balancing and RFECV feature selection methods improves the performance of Random Forest, making it a promising approach for enhancing prediction models.

Keywords: *Random Forest, Balancing Data, Feature Selection, Heart Disease, Disease Prediction*

1. Introduction

It has been determined that heart disease remains the primary cause of mortality on a global scale [1][2]. According to data from the World Health Organization (WHO), in 2019, an estimated 17.9 million individuals perished from heart disease, constituting 32% of the global mortality toll [3]. In Indonesia, data from BPJS Health indicates that as of May 2024, 1.89 million individuals had been diagnosed with heart disease. In addition to its status as the foremost cause of mortality, heart disease constitutes a substantial economic encumbrance on global health systems [4]. According to BPJS Kesehatan, the economic burden is estimated to reach IDR 67.34 trillion [5]. This disease has far-reaching consequences for public health, as it also exerts significant pressure on national productivity. This is particularly salient given that the productive age group is among the most vulnerable to the disease [6]. If not adequately addressed, an increase in heart disease cases can result in greater economic losses due to its impact, which extends from micro to macro scale [7].

One of the strategic steps that can be applied in helping to reduce the impact of heart disease is through the utilization of technology, particularly in the development of data-based prediction models [8]. In the field of early diagnosis, there has been a notable increase in the utilization of machine learning algorithms [9][10][11]. Machine learning, a branch of artificial intelligence, allows computers to identify patterns within data and generate predictions or make decisions without being explicitly programmed [12]. This technology has been applied in various fields, including medicine, to improve the efficiency of diagnosis and treatment [13][14]. The application of machine learning has the potential to enhance the accuracy and efficiency of predicting heart disease, thereby reducing the risk of complications and long-term treatment costs [15].

However, the application of machine learning algorithms as a prediction model faces several challenges, especially with regard to data imbalance between the majority class and the minority class [16][17]. Data imbalance is a phenomenon in which there is an uneven distribution of data points across different classes. This imbalance can cause the model to be biased towards the majority class, leading to a reduction in its ability to accurately identify data points from the minority class. This issue is particularly problematic in machine learning models, where the imbalance can result in a decrease in the model's performance in identifying minority class data points [18]. Furthermore, the extensive utilization of features in the dataset can impede the efficacy of the model, particularly when certain features are found to be irrelevant or redundant [19][20]. Consequently, machine learning algorithms necessitate supplementary methods or techniques to enhance overall performance [21].

Research on heart disease detection has been conducted by several researchers by applying supervised learning-based algorithm models [22]. A review of the extant literature reveals a range of findings regarding the accuracy of these measurements. For instance, a study

examined coronary artery calculation scores to predict coronary heart disease risk employing various methods, including Random Forest, radial basis function neural network (RBFFN), SVM, KNN, and kernel ridge regression (KRR) [23]. A comparative analysis of the five methods reveals that Random Forest demonstrates optimal performance, exhibiting 78.96% accuracy, 93.86% sensitivity, 51.13% specificity, MCC 0.5192, and AUC 0.8375.

Another research focuses on the comparison of Random Forest and SVM methods in diabetes prediction [24]. This study employed data sampling methods such as SMOTE, ENN, and a combination of SMOTE-ENN. The findings indicated that the Random Forest with SMOTE-ENN approach attained 95.8% accuracy, 98.3% sensitivity, and 92.5% specificity. Concurrently, the SVM with SMOTE-NN attained 90% accuracy, 91.1% sensitivity, and 88.5% specificity. These findings suggest that the application of the SMOTE-ENN method significantly enhances the classification performance of diabetic diseases in comparison to the use of SMOTE or ENN separately.

Another research study aims to provide a comparative analysis of various machine learning models in detecting code smells. The study utilizes four distinct datasets: Blob Class, Data Class, Long Parameter List, and Switch Statement [25]. The preprocessing methods employed include the use of SMOTE to address class imbalance and Recursive Feature Elimination (RFE) for feature selection. The models that were evaluated in this study include Extreme Gradient Boosting, AdaBoost, Random Forest, Artificial Neural Network (ANN), and Ensemble Model of Bagging and Boosting Classifiers (EMBBC). The findings indicate that the integration of SMOTE and RFE leads to a substantial enhancement in the classification performance, with the EMBBC model attaining the maximum accuracy of 99.21% on the Blob Class, Data Class, and Long Parameter List datasets. Concurrently, the ANN model attained the maximum accuracy of 92.86% on the Switch Statement dataset. In light of the foregoing, the present research proffers a methodology for the detection of heart disease, employing a hybrid sampling approach that incorporates the SMOTE-ENN method and RFECV feature selection with the Random Forest classification algorithm.

2. Research methods

This section outlines the methodology applied in this study, starting from data collection to evaluation. The research was conducted through a systematic sequence of processes designed to optimize heart disease prediction using machine learning. An overview of the research workflow is presented in Fig. 1, illustrating the key stages involved in the process, including data preprocessing, data balancing, feature selection, classification, and evaluation.

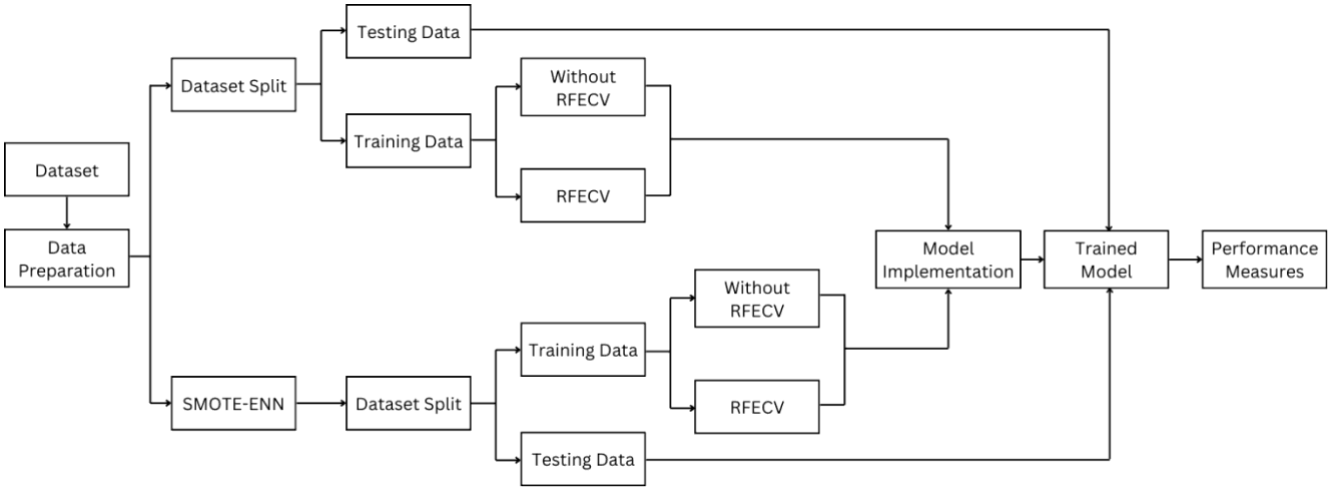

Fig. 1: Research Workflow

2.1. Data collection

In order to assess the efficacy of the proposed algorithm, five groups of datasets were obtained from the UCI Machine Learning Repository. The selection of these datasets was driven by two key factors: their relevance in the context of heart disease prediction and the diversity of their class distributions. The proportion of majority and minority classes varies across each dataset, thereby enabling a more profound examination of the algorithm's efficacy, particularly in scenarios where data imbalance is prevalent. The data set under consideration comprised 1,190 samples drawn from the Cleveland, Hungarian, VA Long Beach, Switzerland, and Statlog datasets. A closer look reveals that the five datasets under consideration share a total of 14 common attributes. Of these, 13 attributes serve as features for analysis, while 1 attribute is a target variable that indicates the diagnosis of heart disease. The details of the heart disease datasets used in this study, including the number of samples and the distribution of minority and majority classes for each dataset, are summarized in Table 1.

Table 1: Dataset Description

No	Attributes	Samples	Minority Sample	Majority Sample
1	Cleveland	303	139	164
2	Hungarian	294	106	188
3	Long Beach VA	200	51	149
4	Switzerland	123	8	115
5	Statlog	270	120	150

2.2. Data preprocessing

The pre-processing stage in this research involves a series of steps aimed at cleaning, transforming, and preparing the data for machine learning algorithms to use. The preliminary stage of the data analysis process involves the purification of data and the standardization of variables. The objective of the data cleaning process is to address issues pertaining to data quality, including missing values and anomalies. Missing values in the dataset are addressed through the implementation of a hot-deck imputation technique, which involves the replacement of missing values in a feature with values extracted from other data elements within the dataset that exhibit similar or most similar characteristics. Subsequent to the cleansing of the data, feature scaling is executed through the utilization of the MinMaxScaler method. This technique involves the standardization of each numeric feature into a range of 0 to 1, thereby ensuring that all features are expressed on a uniform scale. Following the completion of data cleaning and normalization, the dataset is divided into two distinct subsets: training data and testing data. In this study, three different data splitting ratios are employed: 70:30, 75:25, and 80:20. These ratios are used to distribute the data for training and testing purposes.

2.3. Data balancing with SMOTE-ENN

The data balancing technique employed in this research is the hybrid SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbors) method. This method is a combination of oversampling and undersampling techniques, designed to create a more balanced class distribution while improving data quality. The Synthetic Minority Oversampling Technique (SMOTE) is a data mining method that oversamples minority classes by creating new synthetic samples from interpolating values between adjacent minority data points. Meanwhile, Edited Nearest Neighbors (ENN) is an undersampling technique that aims to improve dataset quality by removing majority samples that could potentially cause ambiguity or noise. ENN identifies and removes the majority examples that do not have close enough nearest neighbors from the minority class, thus cleaning the dataset from irrelevant data. SMOTE-ENN integrates the benefits of SMOTE, which involves the addition of synthetic data to enhance representativeness, with the advantages of ENN, which is designed to remove noise from data. In this study, several combinations of SMOTE-ENN parameters were tested to evaluate their impact on model performance. The tested parameter combinations are described in Table 2 below:

Table 2: SMOTE-NN Hyperparameter

Hyperparameter	Description	Value
sampling_strategy	The ratio of the number of minority class samples to the majority class after the oversampling process with SMOTE	1 = 100% 0.95 = 95% 0.9 = 90%
n_neighbors	The number of nearest neighbors used by ENN to remove inconsistent majority data samples	3 5 7

2.4. Feature selection with RFECV

Subsequent to achieving balance within the dataset, the subsequent step involves the implementation of a feature selection process that utilizes Recursive Feature Elimination with Cross Validation (RFECV). This process is employed to identify the most pertinent features that contribute to the prediction of heart disease. The RFECV methodology involves a recursive process of feature elimination. Initially, the training of the classification model is conducted using the complete set of features. Subsequently, features that exhibit minimal contribution to the model's performance are systematically removed. At each iteration, cross-validation is performed to evaluate the remaining feature combinations. This process ensures that the selection process considers not only accuracy on training data, but also generalization to new data. To ensure robustness and to evaluate the impact of validation size, this research applies three different cross-validation fold configurations during RFECV: 3-fold, 5-fold, and 10-fold. These variations allow for an assessment of how different partitioning schemes influence the selection of optimal features and overall model stability. The objective of this process is to enhance the efficiency and accuracy of the classification model by eliminating features that are redundant or have minimal impact on the classification outcomes.

2.5. Classification with random forest

For the purpose of classification, this research employs the Random Forest model as an analytical method. Random Forest is an ensemble algorithm that functions by constructing a multitude of decision trees autonomously. Each decision tree is trained on a random subset of the data, which helps increase variation and reduce the risk of overfitting. Subsequent to the construction of all the decision trees, Random Forest integrates the prediction results from each tree by employing a voting method for classification. This process yields a more stable and accurate final prediction compared to a single decision tree model.

2.6. Model Evaluation Using Confusion Matrix

In the final stage, the performance of the implemented strategy is measured by evaluating it using relevant metrics. This study uses a variety of metrics, including accuracy, precision, recall, and F1-score, to evaluate the model. The evaluation results are subsequently presented in the form of a confusion matrix, thereby providing an overview of the model's performance.

3. Results and discussion

The present study evaluates the performance of the Random Forest model in predicting heart disease under four experimental scenarios: (i) the first method did not include feature selection or data balancing, (ii) the second method employed data balancing using SMOTE-ENN only, (iii) the third method used feature selection using RFECV only, and (iv) the fourth method combined SMOTE-ENN and RFECV. The evaluation is conducted using different data split ratios (70:30, 75:25, and 80:20) and explores various parameter configurations, including SMOTE-ENN's sampling_strategy (1.0, 0.95, 0.90) and n_neighbors (3, 5, 7), as well as the number of cross-validation folds (3, 5, and 10) for RFECV. The performance of the model is evaluated using four standard evaluation metrics: Accuracy, Precision, Recall, and F1-Score. This approach provides a comprehensive understanding of how data balancing and feature selection influence the Random Forest model's predictive capability in heart disease classification. In total, 120 experimental trials were conducted across all scenarios and data split ratios. For clarity and conciseness, the following tables present only the optimal performance achieved for each scenario within its respective data ratio.

Table 3: Optimal Performance Metrics for Each Scenario in 70:30 Data Ratio

Scenario	*sampling_strategy*	*n_neighbors*	*CV Folds*	Performance			
				Accuracy	Precision	Recall	F1-score
1	-	-	-	0.877	0.889	0.889	0.889
2	0.95	5	-	0.949	0.888	0.967	0.926
3	-	-	3	0.882	0.894	0.894	0.894
4	0.90	3	10	0.924	0.891	0.934	0.912

Table 3 presents the performance comparison of all four experimental scenarios under the 70:30 data split configuration, including various parameter settings for SMOTE-ENN and RFECV. The baseline model (Scenario 1), which does not apply any data balancing or feature selection, achieved a moderate F1-Score of 0.889. Scenario 2, which utilizes SMOTE-ENN alone, showed a substantial improvement, reaching the highest F1-Score of 0.926 at a sampling_strategy of 0.95 and n_neighbors of 5. This configuration also resulted in a high Accuracy of 0.949 and a Recall of 0.967, indicating its effectiveness in identifying minority class instances. Scenario 3, involving only RFECV, yielded a slightly better F1-Score of 0.894 (at 3-fold cross-validation) compared to the baseline, demonstrating the advantage of reducing irrelevant features, albeit with less impact than data balancing. Scenario 4, which combines SMOTE-ENN and RFECV, achieved a strong F1-Score of 0.912 with sampling_strategy 0.90, n_neighbors 3, and 10-fold CV, highlighting the benefit of integrating both techniques, although the highest result was still obtained in Scenario 2.

Table 4: Optimal Performance Metrics for Each Scenario in 75:25 Data Ratio

Scenario	*sampling_strategy*	*n_neighbors*	*CV Folds*	Performance			
				Accuracy	Precision	Recall	F1-score
1	-	-	-	0.869	0.880	0.885	0.882
2	0.9	5	-	0.946	0.902	0.949	0.925
3	-	-	5	0.896	0.891	0.894	0.896
4	0.90	5	5	0.946	0.893	0.962	0.926

For the 75:25 data split, Table 4 illustrates that the baseline model (Scenario 1) yielded an F1-Score of 0.882. The application of SMOTE-ENN alone (Scenario 2) once again demonstrated a significant improvement, achieving an F1-Score of 0.925 with a sampling_strategy of 0.90 and n_neighbors set to 5. This configuration also produced a high Accuracy of 0.946 and a strong Recall of 0.949. Scenario 3, which involved feature selection using RFECV only, offered a modest improvement over the baseline, attaining an F1-Score of 0.896 with 5-fold cross-validation. Notably, Scenario 4, which combined both SMOTE-ENN and RFECV, outperformed all other configurations with the highest F1-Score of 0.926, obtained using a sampling_strategy of 0.90, n_neighbors of 5, and 5 CV folds. This result highlights the effectiveness of integrating both data balancing and feature selection, as evidenced by the excellent Recall of 0.962, indicating a highly reliable model for detecting positive cases in this particular data split.

Table 5: Optimal Performance Metrics for Each Scenario in 80:20 Data Ratio

Scenario	*sampling_strategy*	*n_neighbors*	*CV Folds*	Performance			
				Accuracy	Precision	Recall	F1-score
1	-	-	-	0.887	0.930	0.870	0.899
2	0.9	5	-	0.944	0.906	0.935	0.921
3	-	-	10	0.903	0.939	0.891	0.914
4	1	5	5	0.957	0.896	0.984	0.938

For the 80:20 data split, as presented in Table 5, the baseline model (Scenario 1) achieved an F1-Score of 0.899. Scenario 2 (Random Forest with SMOTE-ENN) showed a substantial enhancement in performance, reaching an F1-Score of 0.921 using a sampling_strategy of 0.90 and n_neighbors of 5, accompanied by a high Accuracy of 0.944 and Recall of 0.935. Scenario 3 (Random Forest with RFECV) also improved upon the baseline, with an F1-Score of 0.914 at 10-fold cross-validation, demonstrating RFECV's capability in boosting Precision, which reached 0.939. Most notably, Scenario 4, which integrates SMOTE-ENN and RFECV, delivered the best overall results across all scenarios and data ratios, achieving an F1-Score of 0.938 with sampling_strategy of 1.0, n_neighbors of 5, and 5 CV folds. This optimal configuration also attained the highest Accuracy of 0.957 and an outstanding Recall of 0.984, underscoring the powerful synergy of combining data balancing and feature selection. Such performance is particularly critical in medical diagnostics, where minimizing false negatives is essential for reliable disease detection.

The bar chart below illustrates the average performance of the Random Forest model across four experimental scenarios. ach bar in the chart illustrates the average score of one of the four main evaluation metrics (accuracy, precision, recall, and F1-score) corresponding to a

specific scenario. This visualization offers a concise and interpretable comparison of how different optimization strategies influence the predictive capabilities of the model in heart disease classification.

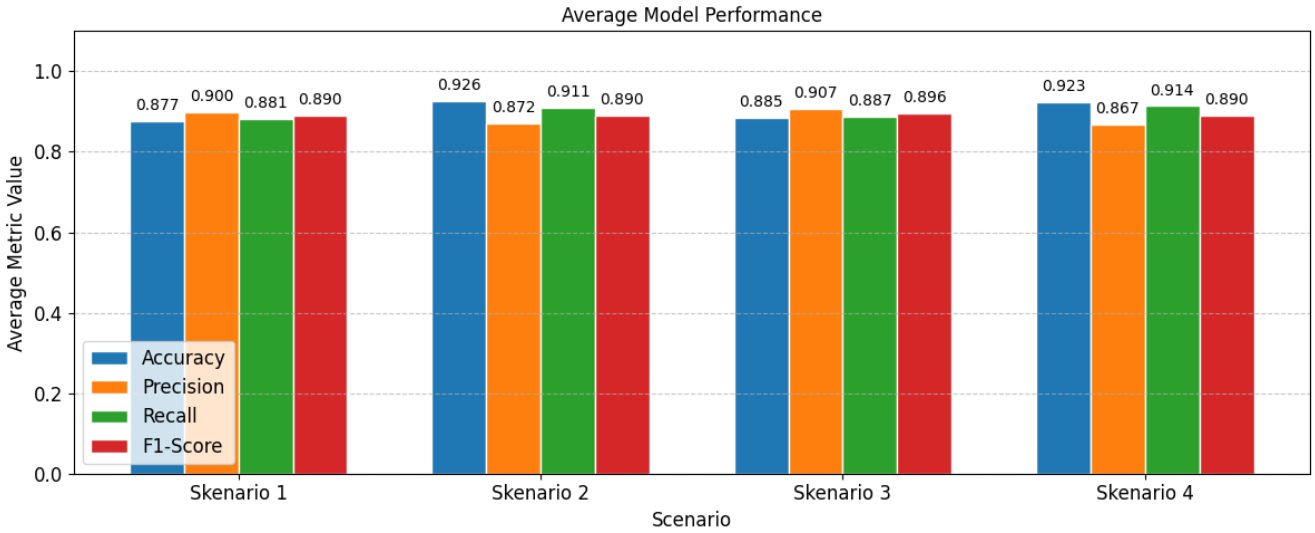

Fig. 2: Average Model Performance in Four Scenarios

The graph in Fig. 2 provides an overall view of the average performance of the Random Forest model on each optimization scenario across test configurations. Scenario 1 (Baseline), which does not apply optimization, demonstrates the lowest average performance with an F1-Score of 0.890, Accuracy of 0.877, Precision of 0.900, and Recall of 0.881. The relatively balanced performance among these metrics at the lowest level confirmed that, in the absence of optimization, the standard Random Forest model experienced challenges in handling the imbalanced and high-dimensional heart disease dataset.

The implementation of SMOTE-ENN in Scenario 2 led to significant enhancements in performance metrics such as Accuracy and Recall, surpassing the performance of the baseline. The mean accuracy increased to 0.926, and the mean recall increased significantly to 0.911, indicating the efficacy of SMOTE-ENN in achieving class balance and enhancing the detection of the minority class (heart disease patients). However, the mean F1-Score remained at 0.890 (equivalent to the baseline), and the mean Precision marginally decreased to 0.872. This indicates a trade-off where aggressive Recall improvement through oversampling may cause the model to predict more positives, including some false positives, thus keeping the F1-Score at the same level despite the increase in true positive case detection.

Scenario 3, which exclusively utilizes RFECV feature selection, demonstrates a more consistent and stable enhancement over the baseline. However, this enhancement is not as pronounced as that observed in Scenario 2 with respect to recall improvement. The mean Accuracy increased to 0.885, the mean Precision reached 0.907 (slightly higher than the baseline), and the mean Recall was 0.887. The mean F1-Score attained 0.896, indicating a consistent and equitable enhancement. These findings substantiate the efficacy of RFECV in identifying the most pertinent subset of features, attenuating noise, and enhancing the model's capacity to make accurate positive predictions. This enhancement in Precision is particularly pronounced, with minimal trade-offs observed in other metrics.

In scenario 4, which integrates both optimization techniques (SMOTE-ENN and RFECV), there is a demonstrable improvement in the average performance of the model when compared to the other scenarios. The mean accuracy achieved an average of 0.923, with the mean recall demonstrating the highest performance among all scenarios at 0.914. While the mean F1-Score remained consistent at 0.890 (equivalent to the baseline and Scenario 2), the mean Precision exhibited a slight decrease to 0.867. However, Scenario 4 demonstrated a notable synergy, particularly in enhancing the detection of true positives (high Recall). The F1-Score, when considered in relation to the Precision, indicates that the model demonstrates a high degree of aggressiveness in detecting positive cases, despite a slight decrease in precision. In the context of heart disease diagnosis, where false negatives carry significantly more severe consequences than false positives, the substantial enhancements in recall and accuracy render this integrated approach optimal for enhancing the generalizability and performance of heart disease prediction models.

4. Conclusion

This study demonstrates the effectiveness of utilizing SMOTE-ENN for data balancing and RFECV for recursive feature elimination to significantly improve the performance of the Random Forest algorithm in predicting heart disease. Through 120 comprehensive experimental trials across varying data split ratios (70:30, 75:25, and 80:20) and diverse parameter configurations for both SMOTE-ENN (sampling_strategy, n_neighbors) and RFECV (CV Folds), the results consistently showed that the proposed hybrid approach achieved substantial gains in model accuracy and recall. Compared to the baseline model and individually optimized models, the combined use of SMOTE-ENN and RFECV consistently yielded higher recall rates, crucial for identifying positive cases. Specifically, the optimal configuration, prominently found with an 80:20 data split, achieved an impressive F1-score of 0.938 and an outstanding recall of 0.984. These results highlight the model's increased sensitivity in identifying positive heart disease cases, which is particularly valuable in clinical decision-making, where false negatives must be minimized.

Acknowledgement

The authors would like to extend special thanks to their academic supervisors for their invaluable guidance throughout the development of this study. The authors would also like to thank their fellow students, as well as everyone who provided moral support, technical insight, or constructive feedback during the research process. Their encouragement and contributions, both direct and indirect, played an important role in the successful completion of this work.

References

[1] A. Singh, H. Mahapatra, A. K. Biswal, M. Mahapatra, D. Singh, and M. Samantaray, "Heart Disease Detection Using Machine Learning Models," *Procedia Comput Sci*, vol. 235, pp. 937–947, 2024, doi: 10.1016/j.procs.2024.04.089.

[2] F. Febby, A. Arjuna, and M. Maryana, "Dukungan Keluarga Berhubungan dengan Kualitas Hidup Pasien Gagal Jantung," *Jurnal Penelitian Perawat Profesional*, vol. 5, no. 2, pp. 691–702, Mar. 2023, doi: 10.37287/jppp.v5i2.1537.

[3] WHO, "Cardiovascular diseases (CVDs)," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[4] M. Wahidin, R. I. Agustiya, and G. Putro, "Beban Penyakit dan Program Pencegahan dan Pengendalian Penyakit Tidak Menular di Indonesia," *Jurnal Epidemiologi Kesehatan Indonesia*, vol. 6, no. 2, Jan. 2023, doi: 10.7454/epidkes.v6i2.6253.

[5] R. Sri Widyastuti, S. Pangarso Wisanggeni, and S. Rejeki, "Beban Ekonomi Penyakit Jantung Rp 67,34 Triliun," Kompas. [Online]. Available: https://www.kompas.id/artikel/beban-ekonomi-penyakit-jantung-rp-674-triliun

[6] Rokom, "Cegah Penyakit Jantung dengan Menerapkan Perilaku CERDIK dan PATUH," Kementerian Kesehatan. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230925/4943963/cegah-penyakit-jantung-dengan-menerapkan-perilaku-cerdik-dan-patuh/#:~:text=Penyakit%20jantung%20masih%20menjadi%20penyebab%20kematian%20nomor%20satu,mortalitasnya%20menyebabkan%20beban%

[7] I. Johanis, I. A. Tedju Hinga, and A. B. Sir, "Faktor Risiko Hipertensi, Merokok dan Usia terhadap Kejadian Penyakit Jantung Koroner pada Pasien di RSUD Prof. Dr. W. Z. Johannes Kupang," *Media Kesehatan Masyarakat*, vol. 2, no. 1, pp. 33–40, Jul. 2020, doi: 10.35508/mkm.v2i1.1954.

[8] F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli, and J. N. Kutz, "Data-driven modeling and learning in science and engineering," *Comptes Rendus. Mécanique*, vol. 347, no. 11, pp. 845–855, Nov. 2019, doi: 10.1016/j.crme.2019.11.009.

[9] S. Asif *et al.*, "Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision," *Archives of Computational Methods in Engineering*, vol. 32, no. 2, pp. 853–883, Mar. 2025, doi: 10.1007/s11831-024-10148-w.

[10] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda.," *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, pp. 8459–8486, 2023, doi: 10.1007/s12652-021-03612-z.

[11] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review.," *Healthcare (Basel)*, vol. 10, no. 3, Mar. 2022, doi: 10.3390/healthcare10030541.

[12] E. Mardiani *et al.*, "Membandingkan Algoritma Data Mining Dengan Tools Orange untuk Social Economy," *Digital Transformation Technology*, vol. 3, pp. 686–693, May 2023, doi: 10.47709/digitech.v3i2.3256.

[13] A. R. Afandi and H. Kurnia, "Revolusi Teknologi: Masa Depan Kecerdasan Buatan (AI) dan Dampaknya Terhadap Masyarakat," *Academy of Social Science and Global Citizenship Journal*, vol. 3, no. 1, pp. 9–13, Jun. 2023, doi: 10.47200/aossagcj.v3i1.1837.

[14] R. R. Chandan *et al.*, "Reviewing the Impact of Machine Learning on Disease Diagnosis and Prognosis: A Comprehensive Analysis," *Open Pain J*, vol. 17, no. 1, May 2024, doi: 10.2174/0118763863291395240516093102.

[15] M. Sharma, J. D. Pandya, R. Thakkar, R. K. Sharma, A. Chopra, and R. S. Tyagi, "Comparative Analysis of Machine Learning Algorithms For Heart Disease Prediction: A Focus On Feature Importance and Model Performance," in *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, IEEE, Dec. 2024, pp. 1433–1437. doi: 10.1109/ICAC2N63387.2024.10894825.

[16] N. V. D. S. S. V. Prasad Raju and P. N. Devi, "A Comparative Analysis of Machine Learning Algorithms for Big Data Applications in Predictive Analytics," *International Journal of Scientific Research and Management (IJSRM)*, vol. 12, no. 10, pp. 1608–1630, Oct. 2024, doi: 10.18535/ijsrm/v12i10.ec09.

[17] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, May 2017, doi: 10.1016/j.neucom.2017.01.026.

[18] W. Nugraha and M. Syarif, "Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost," *Techno.Com*, vol. 22, no. 1, pp. 97–108, Feb. 2023, doi: 10.33633/tc.v22i1.7191.

[19] A. Bailly *et al.*, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Comput Methods Programs Biomed*, vol. 213, p. 106504, Jan. 2022, doi: 10.1016/j.cmpb.2021.106504.

[20] V. Junita and F. A. Bachtiar, "Klasifikasi Aktivitas Manusia menggunakan Algoritme Decision Tree C4.5 dan Information Gain untuk Seleksi Fitur," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 10, pp. 9426–9433, Jan. 2020, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6446

[21] E. Setia Budi, A. Nofriyaldi Chan, P. Priscillia Alda, and M. Arif Fauzi Idris, "RESOLUSI : Rekayasa Teknik Informatika dan Informasi Optimasi Model Machine Learning untuk Klasifikasi dan Prediksi Citra Menggunakan Algoritma Convolutional Neural Network," *Media Online*, vol. 4, no. 5, p. 509, 2024, [Online]. Available: https://djournals.com/resolusi

[22] J. Hamsalatha and S. Renukalatha, "Research on Heart Disease Detection using Machine Learning and Deep Learning Techniques," in *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, IEEE, Oct. 2024, pp. 1–4. doi: 10.1109/SSITCON62437.2024.10797057.

[23] Y. Huang *et al.*, "Using a machine learning-based risk prediction model to analyze the coronary artery calcification score and predict coronary heart disease and risk assessment," *Comput Biol Med*, vol. 151, p. 106297, Dec. 2022, doi: 10.1016/j.compbiomed.2022.106297.

[24] H. Hairani and D. Priyanto, "A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140864.

[25] P. S. Yadav, R. S. Rao, A. Mishra, and M. Gupta, "Ensemble methods with feature selection and data balancing for improved code smells classification performance," *Eng Appl Artif Intell*, vol. 139, p. 109527, Jan. 2025, doi: 10.1016/j.engappai.2024.109527.