Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 2, 2024, pp. 250 – 259

Improving Heart Disease Severity Prediction Using SMOTE for Imbalanced Data

Ayu Hendrati Rahayu*[1], **Ari Sudrajat**[2]
Politeknik TEDC Bandung, Jl. Politeknik – pesantren KM. 2, Cibabat, Cimahi Utara, Kota Cimahi, Jawa Barat 40513
E-mail : ayuhendrati@poltektedc.ac.id[1], arisurajat@poltektedc.ac.id*[2]
**Corresponding author*

Abstract – The heart disease is a prevalent and potentially fatal condition affecting individuals worldwide. In this study, we address the challenge of predicting the severity of heart disease using supervised learning techniques. Leveraging a dataset comprising various demographic and clinical attributes, we propose a solution that employs machine learning models to accurately predict the severity level of heart disease. Among the evaluated models, Random Forest emerges as the top performer, showcasing exceptional precision, recall, accuracy, and F1-score across all severity levels, with an overall accuracy of 98.8%. This highlights the robustness of the Random Forest model in accurately classifying instances across different severity levels. Following closely behind, the KNN algorithm demonstrates commendable performance, achieving an accuracy of 92% and showcasing competitive precision, recall, and F1-score values, particularly for higher severity levels. Despite its notable aspects, XGBoost ranks third among the evaluated models, with an accuracy of 90.4%. While XGBoost excels in certain aspects, such as recall for Level 3 severity, it falls short in overall performance compared to Random Forest and KNN. For future research, exploring ensemble methods that combine the strengths of different algorithms could yield even better classification results, providing avenues for further improvement in predicting the severity of heart disease.

Keywords – Heart Disease, SMOTE, Imbalance Data, Supervised Learning Model, Confusion Matrix.

1. INTRODUCTION

Heart disease, encompassing a range of conditions affecting the heart, remains a leading cause of morbidity and mortality worldwide [1]. Characterized by conditions such as coronary artery disease, arrhythmias, and heart valve problems, heart disease poses significant health challenges and contributes to a substantial economic burden [2]. According to the World Health Organization, heart disease is the number one cause of death globally, accounting for an estimated 17.9 million deaths annually, which represents 31% of all global deaths. In addressing the challenges posed by heart disease, particularly in predicting its severity, machine learning based on supervised learning techniques emerges as a promising solution. By leveraging the wealth of data available on various aspects of heart disease, including patient demographics, medical history, diagnostic tests, and lifestyle factors, supervised learning algorithms can learn intricate patterns and relationships within the data to predict the severity of the disease. Through the utilization of advanced algorithms such as decision trees, support vector machines, or neural networks, coupled with techniques like SMOTE-based oversampling to handle class imbalance, predictive models can be developed to provide early identification of individuals at

higher risk of severe heart disease outcomes [3], [4]. These predictive models not only assist healthcare providers in making informed decisions regarding patient care and treatment strategies but also empower individuals with the knowledge to adopt preventive measures and lifestyle modifications to mitigate the progression of heart disease and reduce its associated morbidity and mortality rates. By integrating machine learning into the realm of cardiovascular healthcare, we stand poised to enhance patient outcomes, reduce healthcare costs, and ultimately alleviate the global burden of heart disease.

Many researchers using supervised learning for heart disease prediction. Such as, Research by Chandrasekhar, et al. [5] focused on enhancing the accuracy of heart disease prediction using machine learning techniques. They explored six algorithms, including random forest, K-nearest neighbor, logistic regression, Naïve Bayes, gradient boosting, and AdaBoost classifier, utilizing datasets from the Cleveland and IEEE Dataport. By employing GridsearchCV and five-fold cross-validation, they optimized model accuracy. In the Cleveland dataset, logistic regression achieved the highest accuracy of 90.16%, while AdaBoost performed best in the IEEE Dataport dataset, reaching 90% accuracy. The study introduced a soft voting ensemble classifier, combining all six algorithms, which further improved accuracy to 93.44% for the Cleveland dataset and 95% for the IEEE Dataport dataset, surpassing the individual performances of logistic regression and AdaBoost. Research by Bhatt, et al. [6] focuses on the development of a machine learning model aimed at improving the diagnosis and prognosis of cardiovascular disease. This model utilizes techniques such as k-modes clustering with Huang starting to enhance classification accuracy. Employing algorithms like random forest, decision tree classifier, multilayer perceptron, and XGBoost, the researchers trained their model on a real-world dataset of 70,000 instances sourced from Kaggle. Through rigorous parameter optimization using GridSearchCV, they achieved impressive accuracy rates ranging from 86.37% to 87.28%. The models also demonstrated high AUC values, indicating robust performance in distinguishing between positive and negative cases of cardiovascular disease. Notably, the multilayer perceptron, particularly with cross-validation, emerged as the top-performing algorithm, boasting an accuracy of 87.28%. Research by Asif, et al. [7] showcases a significant advancement in heart disease prediction through the utilization of machine learning techniques. Their study introduces a comprehensive model that incorporates various preprocessing methods, hyperparameter optimization strategies, and ensemble learning algorithms to accurately predict the presence or absence of heart disease. By merging three datasets from Kaggle and employing the extra tree classifier, data normalization, grid search cross-validation for hyperparameter tuning, and a proper dataset split for training and testing, their approach achieved an impressive accuracy of 98.15%. These findings underscore the potential of their model in early detection, prevention, and management of heart disease, thereby potentially reducing its associated mortality and morbidity rates.

This study proposed herein advances heart disease prediction by employing various supervised learning models, namely K-Nearest Neighbors (KNN), XGBoost, and Random Forest, to address the challenge of imbalanced data. In addressing this issue, the researchers incorporated the Synthetic Minority Over-sampling Technique (SMOTE) to balance both class and attribute distributions within the dataset. By utilizing SMOTE, which synthesizes new instances of the minority class, the researchers aimed to mitigate the effects of data imbalance and improve the overall performance of the predictive models. Through rigorous experimentation and evaluation, the study demonstrated the effectiveness of these techniques in enhancing the accuracy and reliability of heart disease prediction, thereby facilitating early detection and intervention strategies to reduce associated mortality and morbidity rates.

2. RESEARCH METHOD

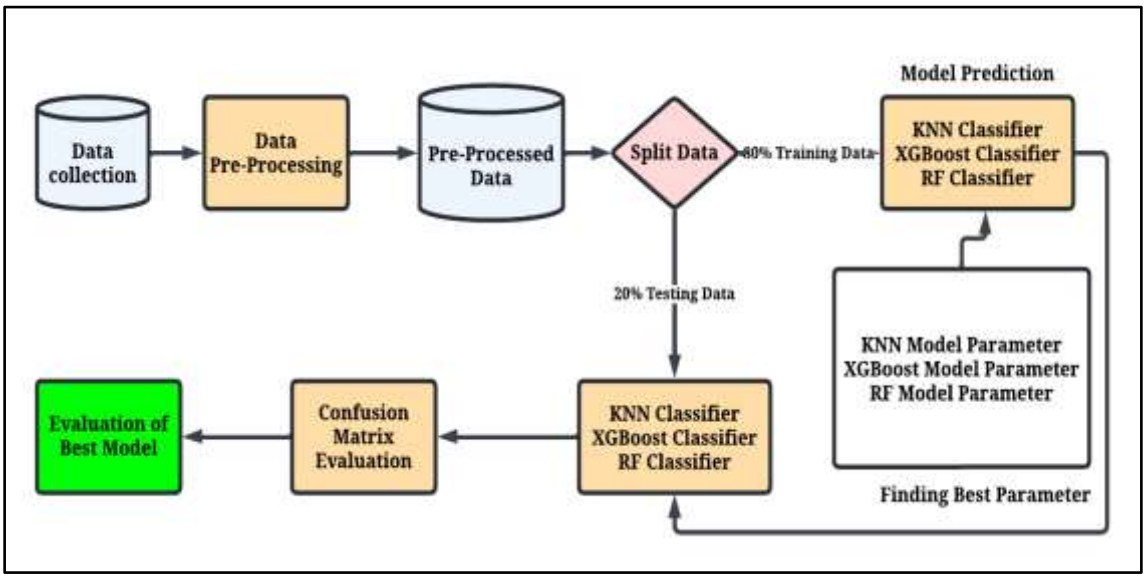


Figure 1. Proposed System Models for Prediction

The proposed system models for heart disease prediction follow a structured methodology encompassing several key stages. Initially, data collection is undertaken from relevant sources to gather comprehensive datasets. Following this, data pre-processing is performed to cleanse and prepare the raw data, resulting in a refined dataset ready for analysis. The pre-processed data is then split into two subsets, with 80% allocated for training and 20% for testing. Three supervised learning models, namely K-Nearest Neighbors (KNN) Classifier, XGBoost Classifier, and Random Forest (RF) Classifier, are employed for model prediction. Each classifier undergoes hyperparameter tuning to identify the optimal parameters, enhancing model performance. Post-training, the models are evaluated using confusion matrices and other evaluation metrics to determine their accuracy and efficacy. The best-performing model is identified through this rigorous evaluation process, ensuring robust and reliable prediction of heart disease severity.

2.1. Data Collection

The dataset used in this study for predicting heart disease comprises 76 attributes, although only 14 key attributes were utilized for the analysis. These attributes include patient demographics and clinical measurements: age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and the diagnosis of heart disease (num), which indicates the degree of vessel narrowing. This selection captures a comprehensive profile of the patient's health, focusing on cardiovascular-related metrics that are crucial for predicting heart disease severity. Based on sample dataset can be seen in Table 1.

Table 1. Sample Datasets and Description

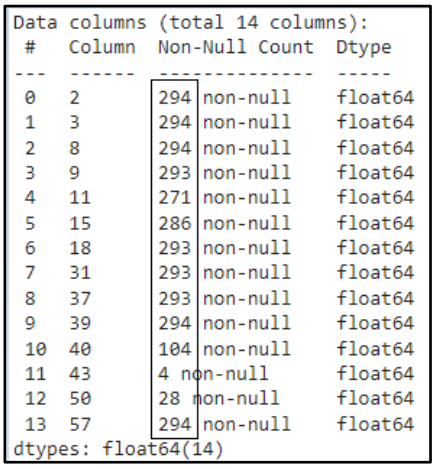
Variable Name	Role	Type	Description	Missing Values
age	Feature	Integer	years	no
sex	Feature	Categorical	-	no
cp	Feature	Categorical	-	no
trestbps	Feature	Integer	Resting blood pressure (on admission to the hospital)	no
chol	Feature	Integer	Serum cholesterol	no
fbs	Feature	Categorical	Fasting blood sugar > 120 mg/dl	no
restecg	Feature	Categorical		no
thalach	Feature	Integer	Maximum heart rate achieved	no
exang	Feature	Categorical	Exercise induced angina	no
oldpeak	Feature	Integer	ST depression induced by exercise relative to rest	no
slope	Feature	Categorical		no
ca	Feature	Integer	Number of major vessels (0-3) colored by fluoroscopy	yes
thal	Feature	Categorical		yes
num	Target	Integer	Diagnosis of heart disease	no

2.2. Pre-Processing

The preprocessing steps proposed for this dataset involve data cleaning to address missing values, wherein missing data points are either removed or imputed with the mean value per attribute. Subsequently, to tackle class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is employed. SMOTE generates synthetic samples for the minority class by interpolating between existing samples in the feature space, thereby balancing the class distribution. By cleaning the data of missing values and applying SMOTE to handle class imbalance, the dataset becomes more robust and suitable for training machine learning models. Pre-processed data can be seen below.

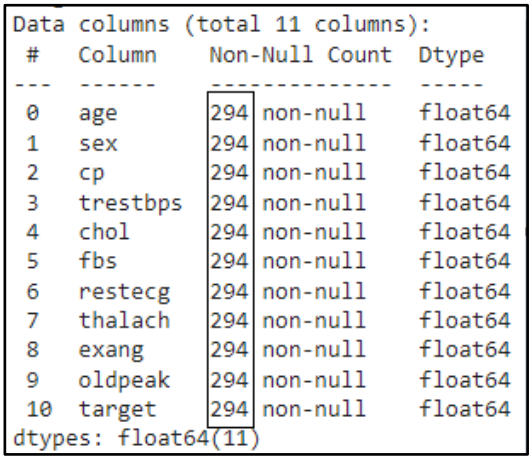
Based on Figure 2 (a), the image illustrates the raw data before undergoing cleaning processes. Notably, numerous missing values are observed across the dataset, indicating inconsistencies and potential data quality issues. However, in Figure 2 (b), after data cleaning, the dataset exhibits consistency, with no missing values detected. Each attribute is now complete, with a total of 294 data points per attribute, ensuring the dataset's integrity and reliability for further analysis. Moving to Figure 2 (c), the visualization depicts the class distribution before applying SMOTE. Here, an evident class imbalance is observed, with one class significantly outnumbering the other. To address this imbalance, Figure 2 (d) presents the class distribution after applying SMOTE. Through SMOTE, synthetic samples are generated for the minority class, resulting in a more balanced distribution between classes.

Based on Figure 2 (a), the image illustrates the raw data before undergoing cleaning processes. Notably, numerous missing values are observed across the dataset, indicating inconsistencies and potential data quality issues. However, in Figure 2 (b), after data cleaning, the dataset exhibits consistency, with no missing values detected. Each attribute is now complete, with a total of 294 data points per attribute, ensuring the dataset's integrity and reliability for further analysis. Moving to Figure 2 (c), the visualization depicts the class distribution before applying SMOTE. Here, an evident class imbalance is observed, with one class significantly outnumbering the other. To address this imbalance, Figure 2 (d) presents the class distribution after applying SMOTE. Through SMOTE, synthetic samples are generated for the minority class, resulting in a more balanced distribution between classes.



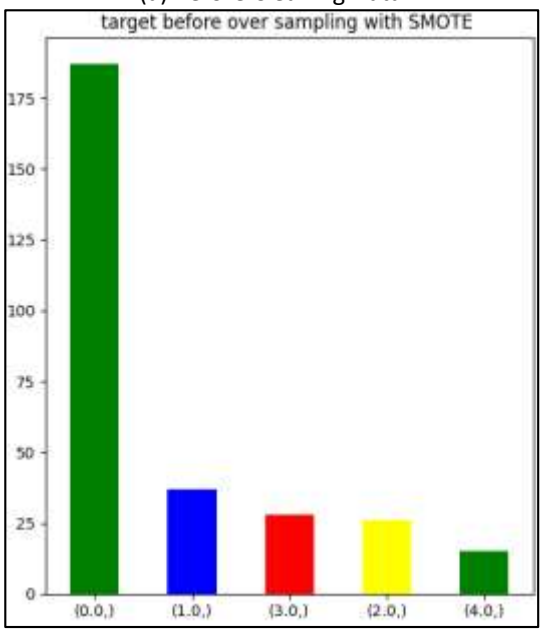
```
Data columns (total 14 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   2       294 non-null    float64
 1   3       294 non-null    float64
 2   8       294 non-null    float64
 3   9       293 non-null    float64
 4   11      271 non-null    float64
 5   15      286 non-null    float64
 6   18      293 non-null    float64
 7   31      293 non-null    float64
 8   37      293 non-null    float64
 9   39      294 non-null    float64
 10  40      104 non-null    float64
 11  43      4 non-null      float64
 12  50      28 non-null     float64
 13  57      294 non-null    float64
dtypes: float64(14)
```

(a) Before Cleaning Data

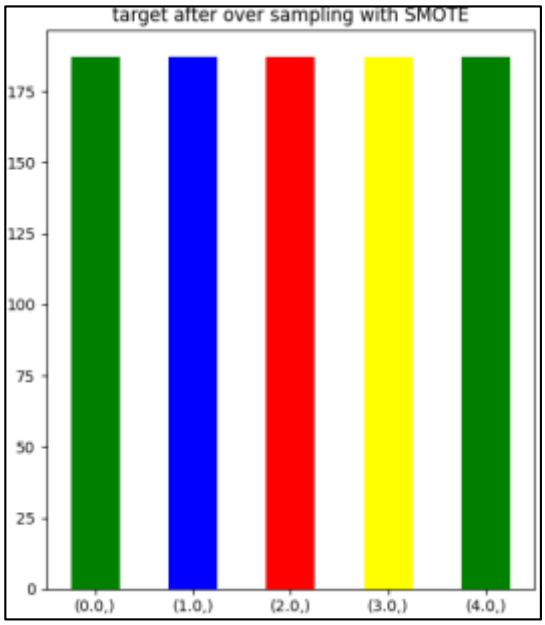


```
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       294 non-null    float64
 1   sex       294 non-null    float64
 2   cp        294 non-null    float64
 3   trestbps  294 non-null    float64
 4   chol      294 non-null    float64
 5   fbs       294 non-null    float64
 6   restecg   294 non-null    float64
 7   thalach   294 non-null    float64
 8   exang     294 non-null    float64
 9   oldpeak   294 non-null    float64
 10  target    294 non-null    float64
dtypes: float64(11)
```

(b) After Cleaning Data



(c) Before Smooting



(d) After Smooting

Figure 2. Pre-processed data

2.3. Supervised Learning Models

Supervised learning models is a type of machine learning algorithm that learns from labeled data, meaning the input data is paired with corresponding output labels [8], [9], [10], [11]. The model is trained on a dataset where both input features and their corresponding correct output labels are provided. During the training process, the model learns the relationship between the input features and the output labels, allowing it to make predictions or decisions when given new, unseen data.

2.3.1 KNN Model

K-Nearest Neighbors (KNN) classifier is a type of supervised learning algorithm used for classification tasks [12], [13]. It is a simple yet powerful algorithm that works based on the principle of similarity. In the KNN algorithm, when given a new, unseen data point, the algorithm identifies the k-nearest data points from the training dataset based on a distance metric (such as Euclidean distance or Manhattan distance). These nearest data points are determined based

on the similarity of their feature values to those of the new data point. Based on euclidean and manhattan distance can be seen in eq (1) and (2).

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{n}(qi - pi)^2} \quad (1)$$

$$Manhattan\ Distance = \sum_{i=1}^{n}|qi - pi| \quad (2)$$

Where, pi and qi are the i^{th} dimensions of point p and q, and n is the number of dimensions.

2.3.2 XGBoost Model

XGBoost, or eXtreme Gradient Boosting, stands as a leading algorithm in the realm of supervised learning, renowned for its exceptional performance in classification and regression tasks [14], [15]. Built upon the gradient boosting framework, XGBoost sequentially combines weak learners, often decision trees, to form a robust predictive model. Its efficacy lies in its ability to address overfitting through L1 and L2 regularization, optimize user-defined objective functions, and prune trees during construction to enhance model simplicity. Moreover, XGBoost offers scalability with support for parallel and distributed computing, facilitating efficient processing of large datasets.

2.3.3 Random Forest Model

Random Forest model stands as a powerful ensemble learning technique renowned for its versatility and robust performance in both classification and regression tasks [16], [17]. Operating by constructing numerous decision trees during training, it amalgamates their outputs to yield a final prediction. Key to its effectiveness is the introduction of randomness at various stages, including feature selection and bootstrapping, which ensures the diversity of individual trees and mitigates overfitting. By leveraging bagging techniques, Random Forest reduces variance and enhances generalization, rendering it less susceptible to noise and outliers in the data. Moreover, its provision of feature importance metrics facilitates insights into the dataset's underlying patterns, aiding in feature selection and interpretation.

2.4. Confusion Matrix

confusion matrix is a tabular representation used in supervised machine learning to evaluate the performance of a classification model. It compares the actual values of the target variable (ground truth) with the predicted values produced by the model [3], [18]. A confusion matrix provides valuable insights into the performance of a classification model, allowing practitioners to assess its accuracy, precision, recall, specificity, and other performance metrics. It serves as the basis for calculating various evaluation metrics, such as accuracy, precision, recall (sensitivity), specificity, F1-score, and area under the ROC curve (AUC-ROC). Based on confusion matrix equation can be seen in eq (3) – (6).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (5)$$

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (6)$$

Where, True Positives (TP), representing instances correctly classified as positive by the model; True Negatives (TN), denoting instances correctly identified as negative; False Positives (FP), or Type I errors, indicating instances inaccurately labeled as positive when they are negative; and False Negatives (FN), or Type II errors, signifying instances erroneously classified as negative when they are positive.

3. RESULTS AND DISCUSSION

The results and discussion section begins with an examination of parameter selection, as outlined in Table 2. This table encapsulates the chosen parameters, providing a comprehensive overview of the model's configuration. Each parameter's significance is scrutinized in light of its impact on model performance, allowing for a detailed analysis of how parameter tuning influences predictive accuracy and generalization. By delving into parameter selection, the section lays the groundwork for elucidating the intricacies of model optimization and shedding light on the factors driving the observed results.

Table 2. Parameter selection per model

Model Classifier	Parameter
KNN Classifier	n_neighbors = 3
XGBoost Classifier	learning_rate=0.1, n_estimators=100, random_state=42
Random Forest Classifier	n_estimators = 100, random_state = 42, max_leaf_nodes = 20, min_samples_split = 15

Following parameter selection, the model undergoes training. Evaluation results of the KNN model are presented in Table 3, providing a comprehensive overview of its performance metrics. Additionally, a visual representation of the confusion matrix based on the KNN model is depicted in Figure 3 (a), offering a clear illustration of the model's classification accuracy and misclassifications across different classes.

Table 3. KNN Evaluation

Severity Level	Accuracy	Precision	Recall	F1-Score
Safe	92%	97%	84%	90%
Level 1		80%	86%	83%
Level 2		92%	95%	93%
Level 3		97%	95%	96%
Level 4		95%	100%	97%

Subsequent to the KNN model evaluation, the XGBoost model undergoes training. Evaluation outcomes of the XGBoost model are displayed in Table 4, enabling a comparative analysis with the KNN results outlined earlier. Additionally, akin to the KNN analysis, the confusion matrix based on the XGBoost model is illustrated in Figure 3 (b), offering insights into the model's classification performance.

Table 4. XGBoost Evaluation

Severity Level	Accuracy	Precision	Recall	F1-Score
Safe	90.4%	92%	89%	91%
Level 1		94%	84%	89%
Level 2		85%	89%	87%
Level 3		88%	99%	93%
Level 4		94%	92%	93%

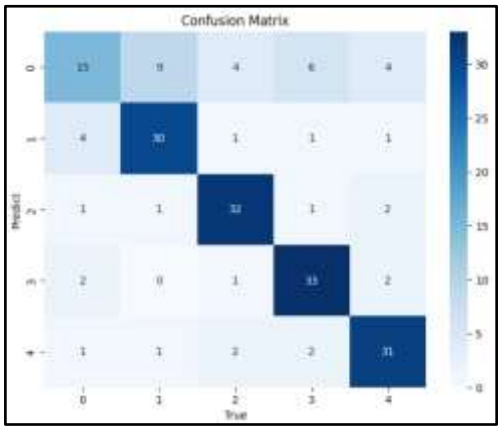
Final evaluation entails training the XGBoost model, building upon the insights gleaned from the KNN assessment. The evaluation results for the XGBoost model are delineated in Table 5, offering a comparative analysis with both the KNN and XGBoost outcomes discussed earlier.

Furthermore, akin to the assessments of KNN and XGBoost, the confusion matrix based on the Random Forest (RF) model is depicted in Figure 3 (c), allowing for a comprehensive examination of classification performance across all three models.

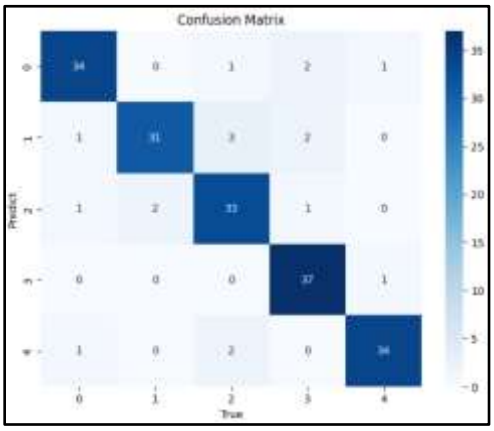
Table 5. Random Forest Evaluation

Severity Level	Accuracy	Precision	Recall	F1-Score
Safe	98.8%	99%	99%	98%
Level 1		94%	95%	96%
Level 2		94%	95%	94%
Level 3		96%	95%	96%
Level 4		100%	100%	100%

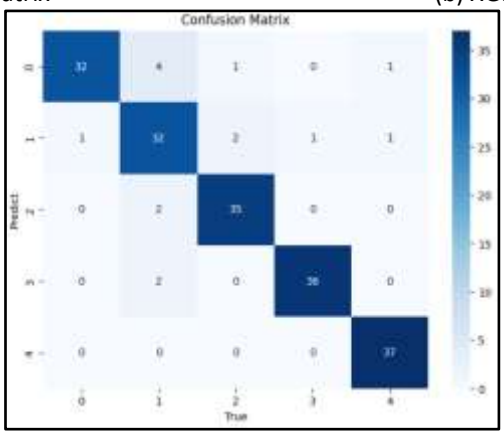
From the evaluation results of the three models conducted, it can be concluded that Random Forest (RF) demonstrates superior performance compared to KNN and XGBoost. RF achieves the highest accuracy across all severity levels, with an overall accuracy reaching 98.8%. Additionally, RF also exhibits excellent precision, recall, and F1-score values for each severity level, indicating its capability in accurately classifying each class. While KNN and XGBoost also demonstrate commendable performance, there are certain severity levels where they do not perform as well as RF. KNN shows decent performance in identifying lower severity levels (Level 1), while XGBoost exhibits better performance in identifying higher severity levels (Level 3). The visualizations of the confusion matrix tables for each respective model can be observed in Figure 3. These visual representations provide a clear and intuitive understanding of the classification performance of each model across different severity levels.



(a) KNN Matrix



(b) XGBoost Matrix



(c) Random Forest Matrix

Figure 3. Table of confusion matrix per model

4. CONCLUSION

Upon evaluating the performance of the KNN, XGBoost, and Random Forest algorithms across various severity levels, it becomes evident that Random Forest (RF) emerges as the top-performing model, followed by KNN and XGBoost. Random Forest exhibits exceptional precision, recall, accuracy, and F1-score across all severity levels, with an overall accuracy of 98.8%. The RF model demonstrates superior precision, recall, and F1-score for each severity level, showcasing its robustness in accurately classifying instances. Following closely behind, the KNN algorithm achieves an accuracy of 92% and demonstrates commendable precision, recall, and F1-score values, particularly for higher severity levels. Despite its competitive performance, XGBoost ranks third among the evaluated models, with an accuracy of 90.4%. While XGBoost excels in certain aspects, such as recall for Level 3 severity, it falls short in overall performance compared to RF and KNN. In conclusion, Random Forest proves to be the most effective algorithm for severity level classification, offering superior performance metrics across precision, recall, accuracy, and F1-score, followed by KNN and XGBoost, respectively.

For future research, exploring ensemble methods that combine the strengths of different algorithms could yield even better classification results. Additionally, investigating the impact of feature engineering techniques and domain-specific knowledge integration on model performance could enhance the accuracy and generalization capabilities of the classifiers. Furthermore, conducting experiments with larger and more diverse datasets could provide deeper insights into the scalability and robustness of the models across various real-world scenarios. Overall, continued research in this area holds the potential to further refine classification models for severity level prediction tasks in healthcare and beyond..

REFERENCES

[1] Y. Wang and D. J. Magliano, “Special Issue: ‘New Trends in Diabetes, Hypertension, and Cardiovascular Diseases,’” *International Journal of Molecular Sciences*, vol. 25, no. 5. Multidisciplinary Digital Publishing Institute (MDPI), Mar. 01, 2024. doi: 10.3390/ijms25052711.

[2] F. Sapna *et al.*, “Advancements in Heart Failure Management: A Comprehensive Narrative Review of Emerging Therapies,” *Cureus*, Oct. 2023, doi: 10.7759/cureus.46486.

[3] I. P. Kamila, C. A. Sari, E. H. Rachmawanto, and N. R. D. Cahyo, “A Good Evaluation Based on Confusion Matrix for Lung Diseases Classification using Convolutional Neural Networks,” *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, p. 0240102, Dec. 2023, doi: 10.26877/asset.v6i1.17330.

[4] N. R. D. Cahyo, C. A. Sari, E. H. Rachmawanto, C. Jatmoko, R. R. A. Al-Jawry, and M. A. Alkhafaji, “A Comparison of Multi Class Support Vector Machine vs Deep Convolutional Neural Network for Brain Tumor Classification,” in *2023 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, Sep. 2023, pp. 358–363. doi: 10.1109/iSemantic59612.2023.10295336.

[5] N. Chandrasekhar and S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” *Processes*, vol. 11, no. 4, Apr. 2023, doi: 10.3390/pr11041210.

[6] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques,” *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.

[7] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization," *Algorithms*, vol. 16, no. 6, Jun. 2023, doi: 10.3390/a16060308.

[8] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: a survey," *PeerJ Comput Sci*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.1045.

[9] M. Liu, M. Jervis, W. Li, and P. Nivlet, "Seismic facies classification using supervised convolutional neural networks and semisupervised generative adversarial networks," *Geophysics*, vol. 85, no. 4, pp. O47–O58, Jul. 2020, doi: 10.1190/geo2019-0627.1.

[10] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding." doi: https://doi.org/10.48550/arXiv.2203.00680.

[11] S. Sakti and B. A. Titalim, "Leveraging the Multilingual Indonesian Ethnic Languages Dataset In Self-Supervised Models for Low-Resource ASR Task," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8. doi: 10.1109/ASRU57964.2023.10389730.

[12] E. H. Rachmawanto *et al.*, "Eggs classification based on egg shell image using k-nearest neighbors classifier," in *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 50–54. doi: 10.1109/iSemantic50169.2020.9234305.

[13] A. Susanto, I. U. W. Mulyono, C. A. Sari, E. H. Rachmawanto, and R. R. Ali, "Javanese Character Recognition Based on K-Nearest Neighbor and Linear Binary Pattern Features," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Sep. 2022, doi: 10.22219/kinetik.v7i3.1491.

[14] A. Farzipour, R. Elmi, and H. Nasiri, "Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods," *Diagnostics*, vol. 13, no. 14, p. 2391, Jul. 2023, doi: 10.3390/diagnostics13142391.

[15] J. Ou *et al.*, "Coupling UAV Hyperspectral and LiDAR Data for Mangrove Classification Using XGBoost in China's Pinglu Canal Estuary," *Forests*, vol. 14, no. 9, p. 1838, Sep. 2023, doi: 10.3390/f14091838.

[16] M. A. Rasyidi, T. Bariyah, Y. I. Riskajaya, and A. D. Septyani, "Classification of handwritten javanese script using random forest algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1308–1315, Jun. 2021, doi: 10.11591/eei.v10i3.3036.

[17] E. H. Rachmawanto, D. R. I. M. Setiadi, N. Rijati, A. Susanto, I. U. W. Mulyono, and H. Rahmalan, "Attribute Selection Analysis for the Random Forest Classification in Unbalanced Diabetes Dataset," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2021, pp. 82–86. doi: 10.1109/iSemantic52711.2021.9573181.

[18] A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowl Based Syst*, vol. 247, Jul. 2022, doi: 10.1016/j.knosys.2022.108651.