Journal of Mongolia International University: Research and Innovation. 1(1)2026

Journal of Mongolia International University: Research and Innovation

Journal homepage: https://jri.miu.edu.mn, journal@miu.edu.mn

Research paper

Comparative Analysis of Classification Algorithms for Cardiovascular Disease Dataset with WEKA

Otgonchuluu Bayarsaikhan

Lecturer. Computer Science Department of Mongolia International University

Corresponding Author: otgonchuluubayarsaikhan@miu.edu.mn

Journal of Mongolia International University: Research and Innovation (2026), Vol. I, No. 1, pp. 71–81

doi: 10.66712/jmri.v1i1.6

Published online: 22 May 22, 2026

Computer Science Department

A B S T R A C T

Predicting Cardiovascular Disease is crucial in healthcare, with machine learning algorithms playing a key role in identifying high-risk patients. Cardiovascular diseases are the leading cause of death globally, highlighting the urgent need for innovative diagnostic tools to predict cardiovascular disease more accurately and efficiently. Through rigorous experimentation across diverse data partitions, we seek optimal hyper-parameter configurations to enhance model accuracy and robustness. This study evaluates the performance of four machine learning classification algorithms namely Multilayer Perceptron, Random Forest, J48, Sequential Minimal Optimization on a cardiovascular disease dataset. Percentage splits from 60% to 95% in 5% increments were employed, with 95% yielding the highest observed accuracy under this experimental setting. Our findings highlight the importance of optimal parameter selection and provide valuable insights into each algorithm, enhancing predictive accuracy in cardiovascular diagnosis.

Keywords: Machine Learning, Classification Algorithms, Percentage Split, Cross-Validation, WEKA

1. Introduction

Cardiovascular diseases (CVD) represent the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually World Health Organization, 2024 [5]. This prevalence underscores the urgency for innovative diagnostic tools and strategies to improve patient outcomes. Early and accurate diagnosis of heart diseases is crucial for effective treatment and prevention.

Recently, machine learning techniques have emerged as promising tools for predicting cardiovascular diseases by analyzing various health parameters. This study focuses on evaluating the performance of four classification algorithms—Multilayer Perceptron (MLP),

Random Forest (RF), J48, and Sequential Minimal Optimization (SMO)—using a CDC dataset. The dataset Kaggle, 2024, employed in this study is derived from the Behavioral Risk Factor Surveillance System (BRFSS) conducted by the Centers for Disease Control and Prevention (CDC) and was accessed through a publicly available Kaggle repository. The dataset includes 18 attributes related to health and lifestyle factors aimed at predicting the presence of heart disease. Accurate assessment of machine learning models necessitates rigorous experimentation with different data splits and cross-validation techniques.

To achieve this, percentage splits ranging from 60% to 95% in 5% increments and performed 5 to 10-fold cross-validation methods were employed in this study. This approach enables a more extensive empirical comparison across multiple training–testing configurations than the commonly used single-split (Gultepe & Rashed, 2019) or fixed cross-

validation strategies (Bashir et al., 2019; Friedman, 2001; Hastie et al., 2009; Weng et al., 2017). By experimenting with a wide range of percentage splits, the aim is to find the optimal balance that maximizes model accuracy and robustness. Similarly, varying the number of folds in cross-validation helps evaluate the model's performance across different subsets of the dataset. Cross-validation is essential for estimating the model's accuracy and robustness, ensuring that the performance metrics are reliable and consistent across various data subsets. Through detailed experimentation, the main objective is to identify the most effective machine learning algorithm and optimal parameter settings for predicting cardiovascular disease.

This study evaluated multiple machine learning algorithms (RF, MLP, J48, SMO) with hyper-parameter tuning on each algorithms on a large scale CDC dataset using both percentage splits (95% yielding optimal results) and 10-fold cross-validation (providing robust performance estimates). Random Forest and MLP consistently achieved high accuracy; however, prior studies El Hamdaoui et al., 2021; Mohsen et al., 2024, emphasize the need for careful hyperparameter tuning to control model complexity and mitigate potential overfitting in high-dimensional datasets.

2. Literature Review

Before exploring the studies related to the classification of cardiovascular disease, it is imperative to establish a clear definition of the term. Cardiovascular disease comprises a range of heart or blood vessel issues, such as coronary artery disease, peripheral artery disease, arrhythmia, valve diseases, and others. These conditions collectively contribute to the spectrum of heart dis- ease Junaid and Kumar, 2020. Numerous studies have been conducted to analyze techniques for predicting cardiovascular disease, with researchers exploring a variety of approaches.

According to World Health Organization, 2024, cardiovascular disease prediction using the KStar, J48, SMO, Bayes Net, and Multilayer Perceptron algorithms on data from the UCI repository. SMO and Bayes Net achieved the highest accuracy, with 89% and 87% respectively, using 10-fold cross-validation. Despite these results, the dataset size of 270 is insufficient for reliable prediction, and overall accuracy is still unsatisfactory.

According to Hastie et al., 2009, coronary artery disease (CAD) as a leading cause of mortality, which can be mitigated through lifestyle changes and medical intervention. This research paper explored various machine learning (ML) models, with and without the use of the synthetic minority oversampling technique (SMOTE), to improve CAD prediction accuracy. The stacking ensemble model, combined with SMOTE and 10-fold cross-validation, achieved the best results, demonstrating the technique's effectiveness. However, even with a relatively large dataset of 6,178 records compared to other studies conducted by (Bashir et al., 2019; Friedman, 2001; Weng et al., 2017; World Health Organization, 2024). These studies concluded that the prediction accuracy of 90.9% is still not sufficient to rely on for CAD prediction.

According to Friedman, 2001, heart disease represents a major global health issue, high- lighting the need for advanced analytical methods to better predict and understand cardiovascular disease (CVD) risk factors. The research, which utilized SPSS and Weka for data analysis, found significant associations between age, body mass index, and CVD. Convolutional neural networks (CNNs) reportedly achieved a prediction accuracy of 98.64% for CVD cases; however, this result was obtained on a limited dataset of 370 records, raising concerns regarding overfitting and gener- alizability. The study also identified increased CVD risk among university students and employees in Saudi Arabia, pointing to the need for better awareness and preventive measures. However, according to the findings, despite the high accuracy of the DL4JMlp-b model, challenges such as long computation times and search convergence issues were encountered. Furthermore, analyz- ing a dataset of 370 records with 10-fold cross-validation is still inadequate for reliable clinical prediction.

According to Bashir et al., 2019, this study aimed to improve prediction accuracy by em- ploying various feature selection techniques and algorithms, including decision trees, logistic re- gression, SVM, naive Bayes, and random forest. The findings indicated that Improved AdaBoost with Random Forest using 10-fold cross-validation achieved the highest accuracy (96.16%) in pre- dicting heart disease. However, even after optimization, the

dataset size of 573 is insufficient for reliable prediction, and despite the impressive results, it is difficult to rely on training data with such a low volume.

According to World Health Organization, 2024, cardiovascular disease (CVD) is the lead- ing cause of death worldwide. A related study in machine learning evaluated the effectiveness of data mining algorithms for predicting cardiovascular disease using the widely recognized Cleve- land heart disease dataset. This dataset comprises 303 patient records with 14 attributes and has been extensively utilized in research, including studies associated with Goldsmiths, University of London, as reported by Weng et al., 2017. Support Vector Machine (SVM), C4.5 Decision Tree, and Naive Bayes algorithms were applied to analyze the data. Feature selection techniques, includ- ing Ranker with InfoGainAttributeEval and SVMAttributeEval, were employed to enhance model performance. While feature selection improved classification accuracy, achieving performance in the range of approximately 85–88%, the results remain limited by the relatively small dataset size, which affects the robustness and reliability of the predictive models for CVD prognosis.

According to Gultepe and Rashed, 2019, heart disease prediction was investigated using data mining techniques on a dataset from the UCI repository, employing the Weka software to compare the Naive Bayes and J48 classifiers. Data was divided into 70% for training and 30% for testing. The Naive Bayes classifier achieved an accuracy of 85.71%, with 78 correctly classified instances, a recall of 0.857, and a precision of 0.859. The J48 classifier initially had an accuracy of 76.92%, with 70 correctly classified instances, a recall of 0.769, and a precision of 0.773. Op- timization using the Bagging method did not improve the Naive Bayes results, but it significantly increased the J48 accuracy to 81.31%. Although Naive Bayes initially showed higher accuracy, J48 outperformed it post-optimization. However, the dataset size of 303 limits the generalizability of cardiovascular disease prediction models.

3. Methodology

3.1. Environment Setup

The experiment has been conducted on a MacBook Pro with M1 Pro chip with 16GB RAM, running on macOS Sequoia (version 15.x).

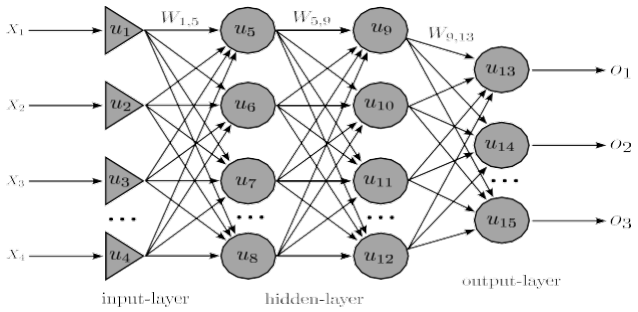


Figure 1. Multilayer Perceptron Wahyunggoro et al., 2013

3.2. WEKA Setup

3.2.1. Background for WEKA

WEKA, short for Waikato Environment for Knowledge Analysis, is a widely respected open-source data mining framework created at the University of Waikato in New Zealand. Known for its extensive range of machine learning algorithms, WEKA offers a variety of data mining tools, including Clustering, Classification, and Association. Its straightforward, adaptable, and user-friendly interfaces make it accessible to users of all experience levels, allowing researchers and practitioners to easily incorporate advanced data mining techniques into their projects Arora, 2012. WEKA version 3.9.6 has been utilized as a research tool.

The percentage split method involves dividing a dataset into two random partitions: a train- ing set and a testing set Saini et al., 2018. For each model evaluation, the model is trained on the training set and tested on the testing set. This approach provides a quick performance estimate, par- ticularly useful for large datasets as it enables efficient training and testing. However, its reliability decreases with smaller datasets due to potential variability in the split.

Cross-validation, on the other hand, divides the dataset into k partitions or folds. A model is trained on k-1 folds and tested on the remaining fold, repeating this process k times so each fold is used as the test set once, resulting in k unique models. The model's performance is then averaged over all k iterations for a more reliable estimate of its

effectiveness. Cross-validation is widely regarded as a commonly used model evaluation technique for both training and testing. However, this approach is computationally intensive, as it requires training multiple models, as noted by Wahyudi and Sumanto, 2024.

3.2.2. Classification Algorithms

According to Dou and Meng, 2023, classification in WEKA is the process of using machine learning algorithms to predict categorical class labels for a dataset. These algorithms learn patterns and relationships within data, enabling them to make accurate predictions on new, unseen instances. In this study, four algorithms were applied to a Cardiovascular Disease Dataset: Multilayer Perceptron (MLP), J48, Random Forest, and SMO.

Multilayer Perceptron (MLP), as shown in Figure 1, is a type of artificial neural network composed of multiple layers of nodes, including an input layer, one or more hidden layers, and an output layer. During training, MLP uses the back-propagation algorithm to adjust the weights of connections between nodes, allowing it to capture complex patterns in the data. MLP is particularly effective for solving nonlinear classification problems, as noted by Arora, 2012; Morariu et al., 2017.

J48, as shown in Figure 2, is the WEKA implementation of the C4.5 decision tree algo- rithm, a popular method for generating decision trees from training data. J48 builds a decision tree by recursively splitting data into subsets based on feature values, aiming to maximize information gain at each split. According to Arora, 2012, a decision tree predicts new instances by traversing from the root node to a leaf node based on the values of the input features.

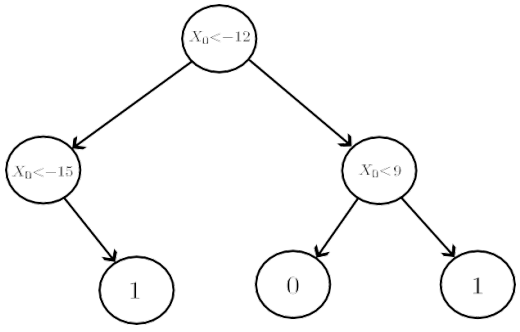


Figure 2. J48

Random Forest (RF), as shown in Figure 3, is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy. It constructs several decision trees, each trained on a random subset of the training data. During prediction, each tree makes a prediction, and final result is determined by majority vote among the trees. According to Mohsen et al., 2024, Random Forest is recognized for its robustness and its capacity to effectively handle complex, high-dimensional datasets.

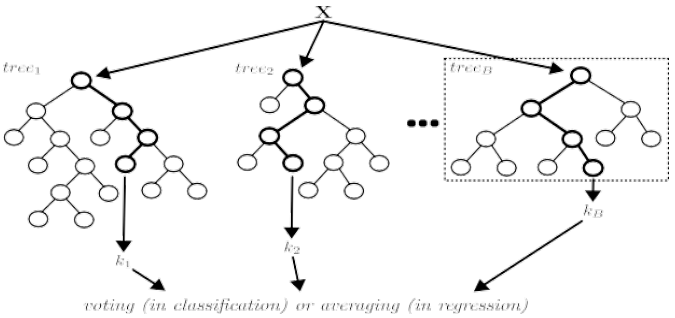


Figure 3. Random Forest Verikas et al., 2016

Sequential Minimal Optimization (SMO), as shown in Figure 4, is a support vector machine (SVM) algorithm used for classification tasks. SVM's aim to find the optimal hyperplane that separates classes while maximizing the margin between the hyperplane and the closest data points. As shown by Arabiat and Altayeb, 2024, SMO provides an efficient solution to the optimization problem in SVM training and is well-suited for large-scale classification tasks.

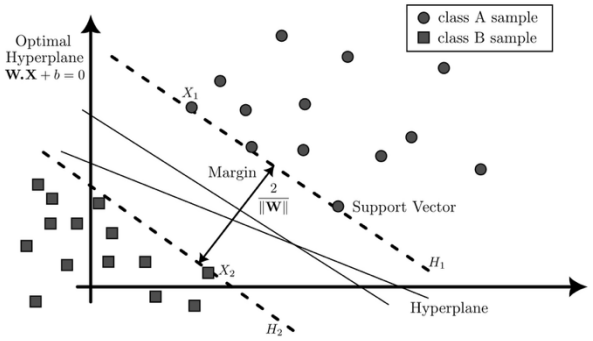


Figure 4. Sequential Minimal Optimization Garcia-Gonzalo et al., 2016

MLP, J48, RandomForest, and SMO Algorithms were experimented on Cross-Validation with Folds of 5-10 and Percentage Split from 60% - 95% with the increment of 5. Algorithm settings were modified as follows in Table 1.

Table 1: Algorithm Settings.

MLP	J48	RandomForest	SMO
Learning Rate:0.5	Confidence: 0.5	Classifiers: True	c: 1.0
momentum:0.3	-	BTR: True	Logistic
GUI: True			Normalize
Epochs: 500	-		RBFKernel: g(0.1)
	-	-	

3.3.Dataset

3.3.1. Data Acquisition

According to the Centers for Disease Control and Prevention (CDC), heart disease stands as a primary cause of mortality in the United States. Major risk factors include high blood pressure, elevated cholesterol levels, and smoking. Additionally, factors such as diabetes, obesity, sedentary lifestyle, and excessive alcohol consumption further contribute to the prevalence of heart disease. In response to this public health concern, machine learning techniques are being employed to analyze CDC data for predicting individuals' susceptibility to heart-related conditions. The dataset Kaggle, 2024, utilized for this research is sourced from the Behavioral Risk Factor Surveillance System (BRFSS), an annual telephone survey conducted by the CDC. The dataset corresponds to the most recent BRFSS release available at the time of data acquisition. This dataset contains 319,795 records, each described by 18 different attributes as shown in Table 1, aimed at understanding the factors related to heart disease condition (HDC) and heart attacks (myocardial infarction). The main outcome is whether a respondent has ever had heart disease, labeled as either "normal" (no heart disease) or "abnormal" (heart disease present). Demographic Variables include gender (male or female), age groups (ranging from 18-24 years up to 80 years or older), and race (such as White, Black, Asian, Hispanic, etc.). These details help us see how heart disease affects different groups of people. Lifestyle and Behavior factors cover body mass index (BMI), smoking habits (whether someone has smoked at least 100 cigarettes in their life), alcohol consumption (more than 14 drinks per week for men or more than 7 for women), and physical activity (whether someone has participated in sports or exercise in the last month). Health Conditions include both physical and mental aspects, such as how many days in the past month the respondent felt physically or mentally unwell. The dataset also includes records about a history of stroke, difficulty walking, and diabetes (including during pregnancy and borderline cases). People were also asked to rate their general health as excellent, very good, good, fair, or poor. Other Health-Related Factors in- clude whether someone has asthma, kidney disease, or skin cancer. It also records average sleep duration per night, as sleep is known to impact heart health.

3.3.2. Preprocessing

After the CDC dataset is acquired, and essential preprocessing steps are executed to ensure data consistency and quality. To begin with, the original CSV file was converted into Attribute Resource Format File (ARFF) as shown in Figure 5, which is the official format of WEKA. Con-secutively, label-encoding was done for the convenience. After loading the data, following filters were employed in order to prepare the dataset for analysis.

- ***RemoveDuplicates*** from unsupervised/instance for removing any duplication from the dataset
- ***InfoGainAttributeEval*** from supervised/attribute with ***Ranker*** was applied to assess feature im-portance; however, no attributes were removed, and all features were retained for model training This includes removing errors and duplicates handling missing values, encoding categorical variables, standardizing the data, and setting the target attribute as nominal class to determine the most impactful features for classification. The initial dataset consisted of 319,795 instances. Following the preprocessing procedures, the dataset was refined and reduced to a total of 301,717 instances. This

reduction resulted primarily from duplicate removal and exclusion of incomplete records.

3.3.3. Model Evaluation

To evaluate the performance of the trained classifiers, we use key metrics including Accuracy, Precision, Recall, and F-Measure. We will explore the definitions, calculations, and interrelationships of these metrics, with TP, TN, FP, and FN representing True Positives, True Negatives, False Positives, and False Negatives, respectively Alanazi and Khamis, 2024.

1. Accuracy, quantifies the proportion of correct predictions among the total predictions made, calculated as *(TP + TN) / (Total instances)*. However, in cases of imbalanced class distribution, accuracy may not serve as a reliable performance metric.

cdcData.arff

```
@relation cdcData

@attribute BMI numeric
@attribute Smoking numeric
@attribute AlcoholDrinking numeric
@attribute Stroke numeric
@attribute PhysicalHealth numeric
@attribute MentalHealth numeric
@attribute DiffWalking numeric
@attribute Gender numeric
@attribute AgeCategory numeric
@attribute Race numeric
@attribute Diabetic numeric
@attribute PhysicalActivity numeric
@attribute GenHealth numeric
@attribute SleepTime numeric
@attribute Asthma numeric
@attribute KidneyDisease numeric
@attribute SkinCancer numeric
@attribute HDC {normal,abnormal}

@data
16.6,1,0,0,3,30,0,0,7,0,1,1,1,5,1,0,1,normal
20.34,0,0,1,0,0,0,0,12,0,0,1,1,7,0,0,0,normal
26.58,1,0,0,20,30,0,1,9,0,1,1,3,8,1,0,0,normal
24.21,0,0,0,0,0,0,0,11,0,0,0,2,6,0,0,1,normal
23.71,0,0,0,28,0,1,0,4,0,0,1,1,8,0,0,0,normal
28.87,1,0,0,6,0,1,0,11,1,0,0,3,12,0,0,0,abnormal
21.63,0,0,0,15,0,0,0,10,0,0,1,3,4,1,0,1,normal
31.64,1,0,0,5,0,1,0,12,0,1,0,2,9,1,0,0,normal
```

Figure 5: CDC file in ARFF format

2. Precision, measures the ratio of true positives *(TP)* to the total predicted positives, calculated as *TP / (TP + FP)*. It evaluates a model's ability to accurately identify true positive instances while minimizing false positive instances.

3. Recall, also referred to as sensitivity or True Positive Rate *(TPR)*, evaluates a model's ability to correctly identify positive instances from all actual positive cases. It is calculated as *TP / (TP + FN)*.

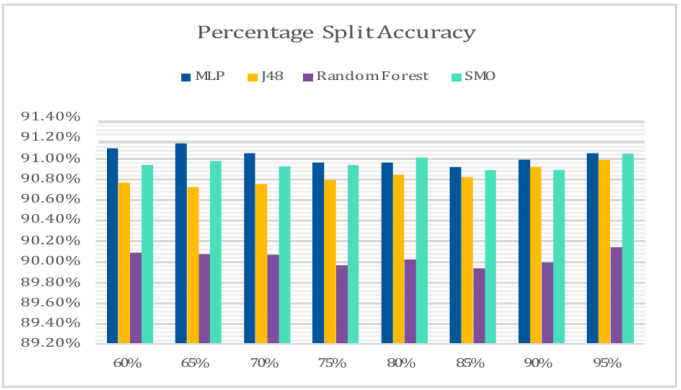
4. F-Measure, the harmonic mean of precision and recall, balances these metrics and is useful for evaluating imbalanced datasets. It is defined as *2 × (Precision × Recall) / (Precision + Recall)*.

This methodology evaluates and optimizes classifier performance, addressing concerns of over-fitting. By systematically assessing each classifier using metrics such as accuracy, precision, recall, and F-Measure across percentage splits from 60% to 95% and cross-validation folds from 5 to 10, more reliable conclusions are drawn. This process facilitates the selection of the most appropriate classification method for CDC data.

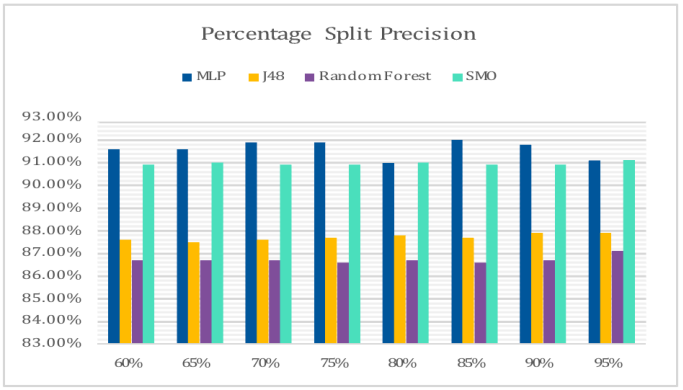
Experimenting with different percentage splits and varying cross-validation folds is essential in evaluating machine learning models for several reasons, and it offers a more robust approach than relying solely on a fixed tenfold cross-validation, as discussed by Bashir et al., 2019; Friedman, 2001; Hastie et al., 2009; Weng et al., 2017 or 70/30 percentage split, as noted by Junaid and Kumar, 2020. Here's why this method is superior:

- *Optimal Data Utilization:* By experimenting with percentage splits from 60% to 95% for training and the remainder for testing, researchers can identify the most effective balance between having sufficient training data to build a robust model and enough testing data to evaluate its performance accurately. Different datasets and models might perform better with different splits.
- *Performance Consistency:* Ensuring the model performs well across various splits demonstrates its consistency and validity.
- *Multiple Scenarios:* Evaluating with different folds (from 5 to 10) covers a range of scenarios, ensuring that the model's performance isn't overestimated or underestimated due to a particular data partition.
- *Robustness and Reliability:* Using different folds tests the model's robustness and reliability, highlighting any potential over-fitting or under-fitting issues that might not be apparent with a single tenfold cross-validation.
- *Fair Evaluation:* Comparing multiple algorithms under varied conditions ensures a fair eval- uation. Some algorithms might perform exceptionally well under certain splits or folds but not others. By using a range of splits and folds, researchers can better understand the strengths and weaknesses of each algorithm.

Prior studies conducted by Aman and Chhillar, 2021; Ansari et al., 2023; Dou and Meng, 2023; El



(a) Percentage Split Accuracy



(b) Percentage Split Precision

Figure 6: Percentage Split Accuracy and Precision

Table 2: Percentage Split Accuracy

Split Ratio	Random Forest	MLP	J48	SMO
60%	90.09%	91.10%	90.77%	90.94%
65%	90.08%	91.14%	90.73%	90.98%
70%	90.07%	91.05%	90.76%	90.93%
75%	89.97%	90.96%	90.79%	90.94%
80%	90.02%	90.96%	90.85%	91.01%
85%	89.94%	90.91%	90.82%	90.89%
90%	90.00%	90.99%	90.92%	90.89%
95%	90.14%	91.05%	90.99%	91.05%

Table 3: Percentage Split Precision

Split Ratio	Random Forest	MLP	J48	SMO
60%	0.867	0.916	0.876	0.909
65%	0.867	0.916	0.875	0.910
70%	0.867	0.919	0.876	0.909
75%	0.866	0.919	0.877	0.909
80%	0.867	0.910	0.878	0.910
85%	0.866	0.920	0.877	0.909
90%	0.867	0.918	0.879	0.909
95%	0.871	0.911	0.879	0.911

Hamdaoui et al., 2021; Gonsalves et al., 2019; Gultepe and Rashed, 2019; Khourdifi and Baha, 2019; Srivastava and Choubey, 2020; Yadav and Pandi, 2019, leveraging cross-validation and percentage split, particularly tenfold and 70/30 percentage split, is crucial for evaluating the performance of machine learning algorithms in medical diagnosis. However, experimenting with percentage splits from 60% to 95% and varying cross-validation folds from 5 to 10 provides a more comprehensive evaluation framework. This approach enhances model robustness, offers detailed performance insights, ensures better generalization, allows fair and balanced algorithm comparison, and optimizes data utilization. These benefits ultimately lead to improved predictive accuracy, aiding in early diagnosis and better clinical decision-making for cardiovascular diseases.

4. Results

4.1. Percentage Split Result

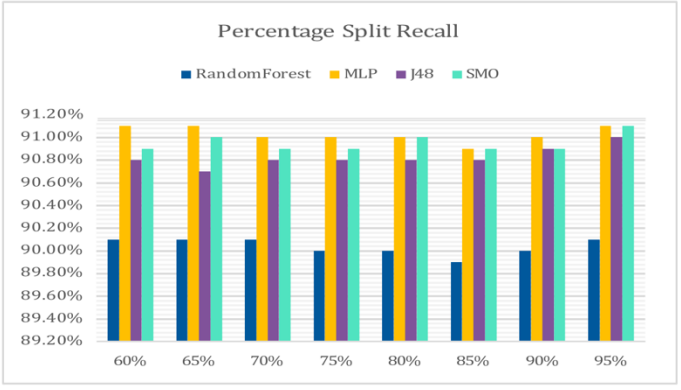
Table 2,3,4,5 and Figure 6a,6b,7a,7b illustrate that among the evaluated configurations, the 95% training split yielded the highest observed accuracy for this dataset. However, this result reflects a more extensive empirical performance under this specific experimental setting rather than guaranteed generalization. At this split ratio, the average

Table 4: Percentage Split Recall

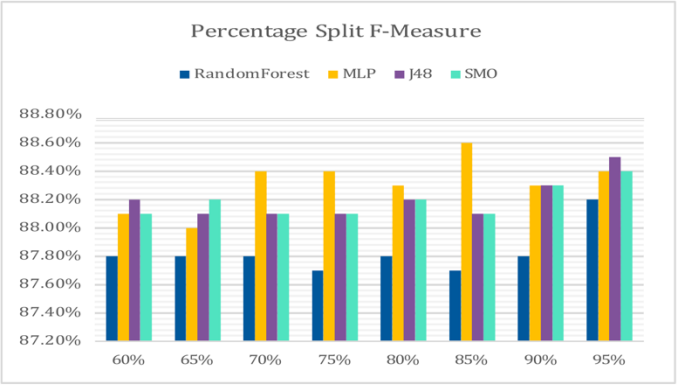
Split Ratio	Random Forest	MLP	J48	SMO
60%	0.901	0.911	0.908	0.909
65%	0.901	0.911	0.907	0.910
70%	0.901	0.910	0.908	0.909
75%	0.900	0.910	0.908	0.909
80%	0.900	0.910	0.908	0.910
85%	0.899	0.909	0.908	0.909
90%	0.900	0.910	0.909	0.909
95%	0.901	0.911	0.910	0.911

Table 5: Percentage Split F-Measure

Split Ratio	Random Forest	MLP	J48	SMO
60%	0.806	0.827	0.808	0.909
65%	0.803	0.840	0.810	0.910
70%	0.800	0.798	0.806	0.909
75%	0.804	0.800	0.812	0.909
80%	0.803	0.910	0.808	0.910
85%	0.799	0.779	0.815	0.909
90%	0.800	0.806	0.805	0.909
95%	0.786	0.911	0.806	0.911



(a) Percentage Split Recall



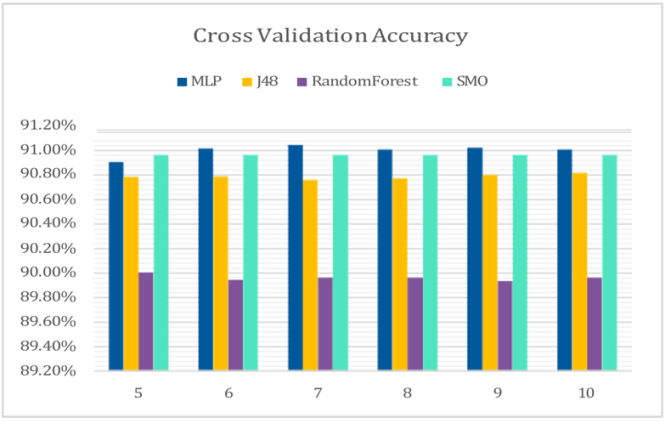
(b) Percentage Split F-Measure

Figure 7: Percentage Split Recall and F-Measure

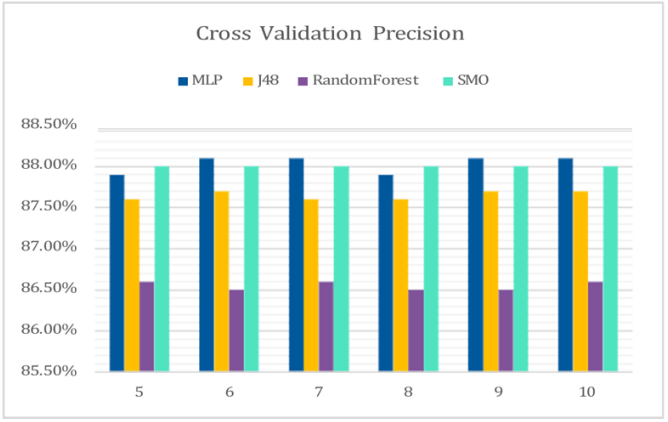
accuracy achieves a high value of 90.81%, indicating the method's effectiveness in making correct predictions. Additionally, the average of precision, recall, and F-measure is also high, with a value of 0.895, indicating balanced performance across various evaluation metrics. Therefore, considering both the accuracy and overall performance metrics, the 95% split yielded the highest observed accuracy in our experiments. However, this does not necessarily imply superior generalization compared to cross-validation on the given big CDC dataset.

4.2. Cross Validation Result

Table 6,7,8,9 and Figure 8a,8b,9a,9b illustrate that the 10-fold cross-validation method demonstrate the highest performance metrics. The average accuracy achieved with 10-fold cross- validation is 90.69%. Additionally, the combined average of precision, recall, and F-measure is 0.888. These metrics indicate that 10-fold cross-validation provides the most reliable and robust estimates of model performance. Therefore, we conclude that 10-fold cross-validation is the most optimal cross-validation method for evaluating the performance of the machine learning algorithms in the given big CDC dataset.



(a) Cross Validation Accuracy



(b) Cross Validation Precision

Figure 8: Cross Validation Accuracy and Precision

Table 6: Cross Validation Accuracy

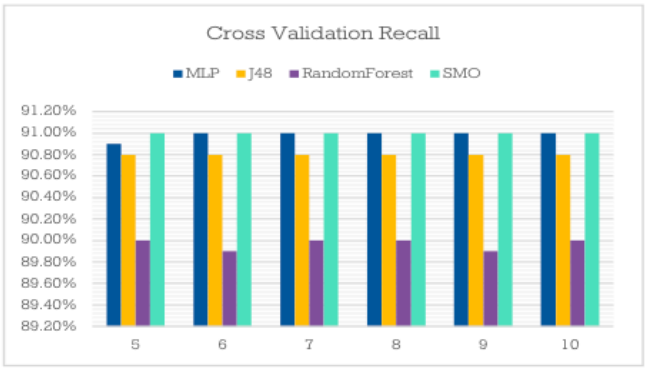
CV	Random Forest	MLP	J48	SMO
5 Folds	90.01%	90.91%	90.79%	90.96%
6 Folds	89.94%	91.02%	90.79%	90.96%
7 Folds	89.96%	91.05%	90.76%	90.96%
8 Folds	89.96%	91.01%	90.78%	90.96%
9 Folds	89.94%	91.03%	90.80%	90.96%
10 Folds	89.96%	91.01%	90.82%	90.96%

Table 7: Cross Validation Precision

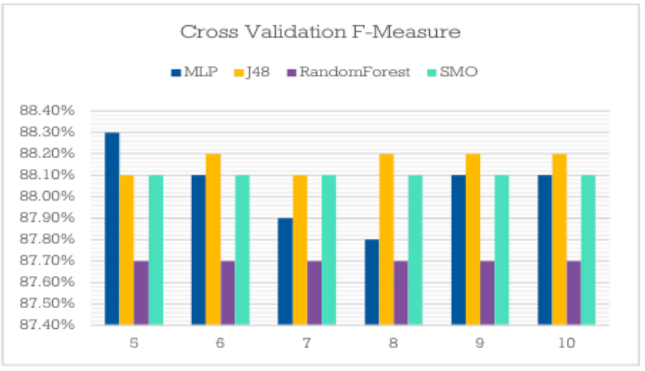
CV	Random Forest	MLP	J48	SMO
5 Folds	0.866	0.879	0.876	0.880
6 Folds	0.865	0.881	0.877	0.880
7 Folds	0.866	0.881	0.876	0.880
8 Folds	0.865	0.879	0.876	0.880
9 Folds	0.865	0.881	0.877	0.880
10 Folds	0.866	0.881	0.877	0.880

5. Conclusion

For the given big CDC dataset, the 95% split yielded the highest observed accuracy under this experimental setting though it may reduce the reliability of performance estimation due to a smaller test set. However, for cross-validation methods, the 10-fold cross-validation provides the most reliable and robust performance estimates, ensuring a thorough evaluation of the machine learning algorithms. Each method has its strengths, and the choice between them may depend on the specific requirements and constraints of the model evaluation process. Random Forest consistently demonstrates high accuracy across various split ratios, indicating its consistency in capturing complex patterns in the data; however, overfitting may remain a concern in general—particularly in high-dimensional datasets. MLP achieves competitive precision and F-measure scores, with averages ranging from 0.906 to 0.920, suggesting its ability to generalize well to unseen data. Still, it requires careful hyper-parameter tuning to mitigate over-fitting, as indicated by the performance across different split ratios. J48 shows slightly lower performance compared to Random Forest and MLP, with average accuracy ranging from 90.73% to 90.92%. This suggests its limitations in capturing complex relationships within the data, particularly in high-dimensional datasets. Similarly, SMO maintains stable performance across split ratios, with average accuracy ranging from 90.89% to 91.05%. However, its computational cost and sensitivity to kernel functions could pose challenges, especially in large-scale applications. In summary,



(a) Cross Validation Recall



(b) Cross Validation F-Measure

Figure 9: Cross Validation Recall and F-Measure

CV	Random Forest	MLP	J48	SMO
5 Folds	0.900	0.909	0.908	0.910
6 Folds	0.899	0.910	0.908	0.910
7 Folds	0.900	0.910	0.908	0.910
8 Folds	0.900	0.910	0.908	0.910
9 Folds	0.899	0.910	0.908	0.910
10 Folds	0.900	0.910	0.908	0.910

Table 9: Cross Validation F-Measure

CV	Random Forest	MLP	J48	SMO
5 Folds	0.804	0.807	0.810	0.881
6 Folds	0.805	0.824	0.809	0.881
7 Folds	0.802	0.841	0.809	0.881
8 Folds	0.805	0.843	0.808	0.881
9 Folds	0.804	0.824	0.808	0.881
10 Folds	0.803	0.822	0.808	0.881

while the percentage split method provides a convenient way to evaluate machine learning models, its reliability diminishes with smaller datasets. Understanding the strengths and weaknesses of each algorithm, as demonstrated by their performance metrics, is crucial for selecting the most appropriate model for a given dataset. Moreover, techniques like cross-validation can complement the percentage split method by providing more robust performance estimates and mitigating the impact of dataset variability on model evaluation. A key limitation of this study was the lack of socioeconomic variables, particularly income status, which could provide deeper insights into cardiovascular disease outcomes. Future research will be aimed towards improving prediction accuracy through the application of advanced techniques such as WEKA's ensemble techniques, nested cross-validation, and incorporating different data sources and the combination of various models within WEKA. It is crucial to thoroughly evaluate how well these models perform across various demographic groups with different datasets.

References

Alanazi, S. M., & Khamis, G. S. M. (2024). Optimizing machine learning classifiers for cardio- vascular prediction. *Engineering, Technology & Applied Science Research*, *14*(1), 12911– 12917. https://doi.org/10.48084/etasr.6684

Aman & Chhillar, R. S. (2021). Analyzing predictive algorithms for cardiovascular disease using weka. *International Journal of Advanced Computer Science*, *12*(8). https://doi.org/10.145 69/IJACSA.2021.0120817

Ansari, G. A., Bhat, S. S., Ansari, M. D., Ahmad, S., & Nazeer, J. (2023). Performance evaluation of machine learning techniques for heart disease prediction. *Computational and Mathemat- ical Methods in Medicine*, *2023*, 8191261. https://doi.org/10.1155/2023/8191261

Arabiat, A., & Altayeb, M. (2024). Assessing data mining tools in traffic congestion prediction. *Indonesian Journal of Electrical Engineering*, *34*(2), 1295–1303. https://doi.org/10.11591/ijeecs.v34.i2.pp1295-1303

Arora, R. (2012). Comparative analysis of classification algorithms on different datasets using weka. *International Journal of Computer Applications*, *54*(13).

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. *2019 16th International Bhurban Confer- ence on Applied Sciences and Technology (IBCAST)*, 619–623. https://doi.org/10.1109 /IBCAST.2019.8667106

Dou, Y., & Meng, W. (2023). Comparative analysis of weka-based classification algorithms on medical datasets. *Technology and Health Care*, *31*(S1), 397–408. https://doi.org/10.3233/t hc-236034

El Hamdaoui, H., Boujraf, S., Chaoui, N. E. H., Alami, B., & Maaroufi, M. (2021). Improving heart disease prediction using random forest and adaboost. *International Journal of Online Engineering*, *17*(11), 61. https://doi.org/10.3991/ijoe.v17i11.24781

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Garcia-Gonzalo, E., Fernandez-Muniz, Z., & Garcia Nieto, P. J. (2016). Hard-rock stability analy- sis using learning classifiers. *Materials*, *9*(7), 531. https://doi.org/10.3390/ma9070531

Gonsalves, A. H., Thabtah, F., Mohammad, R., & Singh, G. (2019). Prediction of coronary heart disease using machine learning. *International Conference on Deep Learning Technologies*, 51–56. https://doi.org/10.1145/3342999.3343015

Gultepe, Y., & Rashed, S. (2019). The use of data mining techniques in heart disease prediction.

International Journal of Computer Science and Mobile Computing, *8*(4), 136–141. https://doi.org/10.47760/IJCSMC

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Junaid, M. J. A., & Kumar, R. (2020). Data science and its application in heart disease prediction.*2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 396–400. https://doi.org/10.1109/ICIEM48762.2020.916 0056

Kaggle. (2024). Dangerous heartbeat dataset (dhd) [[Accessed November 18, 2024]]. https://www .kaggle.com/code/abosalem/dangerous-heartbeat-dataset-dhd-acc-92

Khourdifi, Y., & Baha, M. (2019). Heart disease prediction using optimization algorithms. *Inter- national Journal of Intelligent Systems*, *12*(1).

Mohsen, A. A., Kharroubi, N., Alrashahy, T., & Noaman, S. (2024). Classification and diagnosis of heart disease using machine learning. *Research Square*. https://doi.org/10.21203/rs.3.rs -3985932/v2

Morariu, D., Cretulescu, R., & Breazu, M. (2017). The weka multilayer perceptron classifier. *In- ternational Journal of Advanced Statistics and IT&C*, *7*(1).