Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

Cardiovascular Disease Risk Prediction Using Random Forest, RFECV Feature Selection, and SHAP with Multisource Clinical Data Integration

Dea Fania*[1], Indra Waspada[2], Helmie Arif Wibawa[3]

[1]Master Program of Information Systems, Universitas Diponegoro, Indonesia
[2,3]Department of Informatics, Universitas Diponegoro, Indonesia

Email: [1]deafnaa@gmail.com

Received : Feb 21, 2026; Revised : Feb 26, 2026; Accepted : Feb 26, 2026; Published : Feb 26, 2026

Abstract

Cardiovascular disease (CVD) remains one of the leading causes of mortality in Indonesia, highlighting the urgent need for effective preventive strategies, including the development of risk prediction systems based on population health data. A major challenge in developing CVD prediction models is the limited availability of local medical data that adequately represent the Indonesian population. This study aims to develop a CVD risk prediction model using the Random Forest algorithm by integrating two data sources: private clinical data from cardiology outpatients at RSUD M. Yunus Bengkulu and a publicly available dataset. Data integration was conducted to address the limited size of private data and to improve model performance. The research was conducted through three experimental settings. Shapley Additive Explanations (SHAP) were employed to analyze the contribution of each feature, while Recursive Feature Elimination with Cross-Validation (RFECV) was applied for feature selection. The results indicate that Scenario 3 in the Experiment on Data Integration achieved the best performance, with an accuracy of 73.57%, recall of 81.44%, and F1-score of 77.06%. SHAP analysis identified blood pressure and age as the most influential predictors of CVD risk. These findings demonstrate that integrating limited private data with public datasets can significantly improve model performance while providing clinically interpretable insights, particularly in settings with constrained local data availability.

Keywords : *Cardiovascular Disease, Random Forest, SHAP, Data Integration, Risk Prediction System*

1. INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death worldwide, including in Indonesia, with major risk factors such as unhealthy lifestyles and physiological conditions like hypertension and diabetes [1], [2], [3], [4]. The World Health Organization reports approximately 17.9 million deaths each year due to CVD, of which 85% are caused by heart attacks and stroke [5]. In Indonesia, 877,531 heart disease cases have been recorded based on physician diagnosis, including 6,571 cases in Bengkulu Province [6]. This condition highlights the importance of early CVD detection as a preventive effort to reduce mortality rates [7].

Various studies have developed CVD prediction models based on machine learning (ML) [3], [8], [9], [10], [11], [12]. In the healthcare domain, machine learning (ML) is widely applied for classification and prediction tasks to analyze medical data, identify patterns, and generate accurate predictions of various health conditions [7], [11], [13], [14], [15], [16]. One of the most frequently used ML algorithms is Random Forest [17], [18], [19], [20]. The Random Forest (RF) algorithm has notable advantages in handling complex and heterogeneous data. It operates by randomly sampling the data and aggregating the results from multiple decision trees, forming an ensemble model. Through this mechanism, RF is

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

able to produce more accurate predictions while reducing the risk of overfitting compared to many other machine learning methods [10], [21], [22].

However, the development of CVD prediction models in Indonesia is still constrained by the limited availability of local clinical data, as most studies rely on foreign public datasets and rarely utilize data from the Indonesian population [1], [2], [3], [10], [21], [22], [23], [24]. In addition, research that integrates public datasets with private datasets while applying Recursive Feature Elimination Cross-Validation (RFECV) for feature selection [25], [26], [27] and providing model interpretability through Shapley Additive Explanations (SHAP) [28], [29], [30] under limited data conditions remains very limited. This situation may produce prediction models that are less representative of Indonesian population characteristics and that have suboptimal generalization and clinical interpretability.

Based on these limitations, this study aims to develop a CVD risk prediction model through multisource data integration by combining local Indonesian clinical data (private data) obtained from RSUD M. Yunus, Bengkulu City, with a dataset downloaded from Kaggle (public data). The model is developed using the RF algorithm with RFECV based feature selection and SHAP based interpretability analysis to improve predictive performance and enhance the clarity of model interpretation under limited data conditions.The proposed approach is evaluated through three experimental scenarios: analyzing the effect of reducing the number of features, comparing model performance between public and private data, and evaluating model performance on the integrated public and private dataset. This approach is expected to produce a more accurate CVD risk prediction model while also providing explanations that are easier to understand in a clinical context.

This study offers several key contributions. First, it proposes a multisource data integration strategy by combining limited local clinical data with a publicly available dataset to address data scarcity and improve model robustness. Second, it applies a RF model with feature selection based on RFECV within the context of integrated data to identify the most relevant predictors. Third, it incorporates SHAP based interpretability analysis to enhance transparency and increase trust in the model's predictions. Overall, this study provides a practical framework for developing an accurate and clinically interpretable CVD risk prediction model, particularly in settings with limited local data availabilit.

2. METHOD

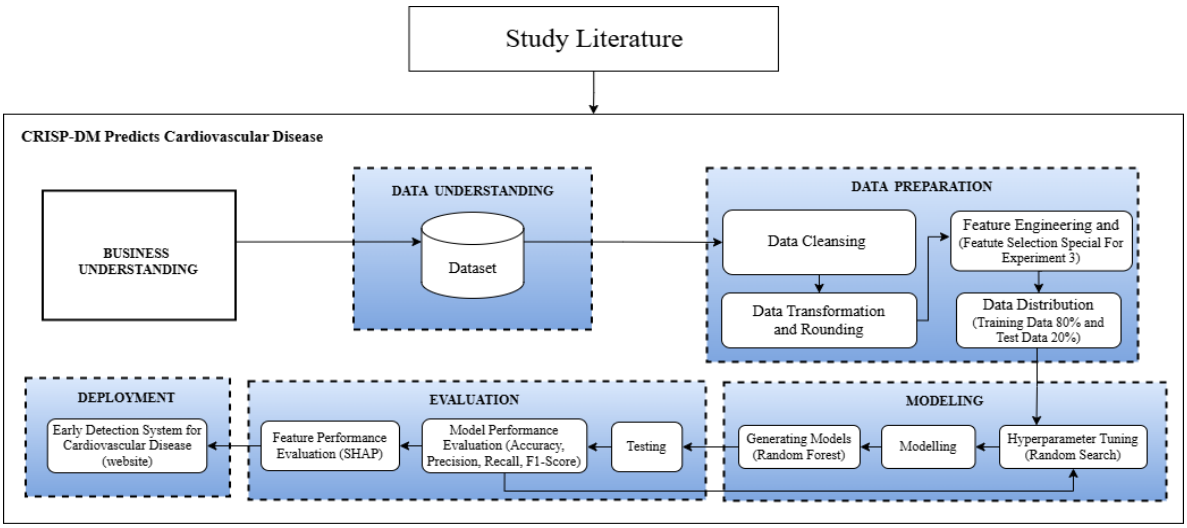

Figure 1. CRISP-DM Research Workflow

This study employs a predictive quantitative approach that focuses on improving the quality and performance of a CVD prediction model derived from a limited local dataset (private data) by

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

incorporating a dataset from another source (public data). The research follows the CRISP-DM methodology as a systematic research framework for developing the CVD model. A visualization of the research workflow is presented in Figure 1.

2.1. Study Literature

This stage aims to review previous studies relevant to the development of CVD risk prediction models. Research on CVD risk prediction has been widely conducted using various machine learning (ML) algorithm approaches, and among all ML algorithms, Random Forest (RF) has proven to be the most stable and accurate for heart disease classification [7], [10], [18], [21], [22], [31]. However, most of these studies rely on foreign public datasets, while studies using local Indonesian CVD datasets remain very limited. This situation poses a challenge in developing prediction models that accurately represent the health conditions and characteristics of the Indonesian population. Therefore, the objective of this study is to integrate a local dataset with a dataset from another source to address the limited size of private data and to improve the performance of the CVD prediction model.

2.2. Business Understanding

The business understanding stage in this study begins by formulating the main objective of the research, which is to improve the quality and performance of a CVD prediction model derived from a limited private dataset by incorporating a dataset from another source and by analyzing the effect of differences in the number of features between the private dataset and the dataset from the other source.

2.3. Data Understanding

In the data understanding stage, an initial exploration was conducted on both the private and public datasets to identify the data used to train the model. The public dataset consists of CVD medical records containing 70,000 entries with 12 attributes [32]. Meanwhile, the private dataset is a local dataset collected directly from the cardiology outpatient medical records at RSUD M. Yunus, Bengkulu City, comprising 250 records with attributes aligned to the public dataset. However, one challenge in collecting the private data is that the available attributes at RSUD M. Yunus, Bengkulu City are incomplete, with only 10 of the 12 required attributes provided.

Table 1. Dataset Description

Attribute	Variable	Unit
Age	Age	Day
Height	Height	Cm
Weight	Weight	Kg
Gender	Gender	0: Female, 1: Male
Systolic blood pressure	Ap_hi	mmHg
Diastolic blood pressure	Ap_lo	mmHg
Cholesterol	Cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Glu	1: normal, 2: above normal, 3: well above normal
Smoking	Smoke	0: No, 1: Yes
Alcohol intake	Alco	0: No, 1: Yes
Physical activity	Active	0: No, 1: Yes
Cardiovascular disease	Cardio	0: No, 1: Yes
Body Mass Index	bmi	kg/m^2
Pulse	Pulse	Beat per minute
Ratio	ap_hi_ap_lo_ratio	Rasio
Blood Pressure Category	bp_category	0: No Hypertension, 1: Hypertension

This study also applies feature engineering by adding four new attributes related to the original data to enrich data variation and improve model evaluation performance [7], [13], [33]. The added attributes include BMI, Pulse (pulse pressure), Ratio (blood pressure ratio), and Blood Pressure Category (hypertensive/non-hypertensive), resulting in a total of 16 attributes, including the label, used in this study. A complete visualization of the dataset is presented in Table 1.

2.4. Data Preparation

The data preparation stage is the process conducted before model training begins. At this stage, the data are prepared to ensure they are clean, relevant, and ready to be processed optimally during modeling. Data preparation starts with removing rows that contain identical values across all columns (duplicate data) to prevent bias in the model. Next, rows containing missing or null values are removed so that the learning process is not affected by incomplete data. Data cleaning also includes eliminating anomalous values that are medically implausible. This step ensures that the data used are realistic and reflect clinical conditions.

The next stage involves data transformation and rounding, which includes converting categorical variables into numerical form and simplifying values to reduce the risk of overfitting. Feature engineering is then performed by adding four derived attributes BMI, pulse pressure, systolic diastolic ratio, and blood pressure category to enrich the research dataset. The final step is splitting the data into 80% training data and 20% testing data across all experimental scenarios. The number of records before and after preprocessing in this study is presented in Table 2.

Table 2. Complete Research Data

Data	Data Source	Before Preprocessing	After Preprocessing
Public Data	Website Kaggle	70.000 Data	60.125 Data
Private Data	RSUD M.Yunus Kota Bengkulu	250 Data	159 Data

2.5. Modelling

The algorithm used in this study is Random Forest because it has strong capability in handling complex and diverse data, such as medical data that involve many attributes, including lifestyle and physiological factors. The modeling process in this research is conducted through several experiments to test and compare the performance of the prediction model under ideal data conditions (complete data) and local data conditions with limited attributes. In implementation, this research utilized Google Colab as the development environment, employing NumPy version 1.26.4 and Scikit-Learn version 1.5.1, executed on a Core i7 laptop processor.

2.5.1. Hyperparameter Tuning

In the hyperparameter tuning process, the Random Search (RS) method is used to obtain the optimal combination of Random Forest parameters, with a batch size setting of 42 and a total of 10 search iterations. The selection of 10 iterations in the Random Search procedure was intended to provide sufficient hyperparameter exploration while avoiding excessive computational complexity, particularly given the limited size of the private dataset. A batch size of 42 was defined to maintain training stability and ensure a balanced learning proportion during each stage of the hyperparameter tuning process. In practice, RS evaluates various predefined hyperparameter values, including n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap, criterion, and class_weight. After the search process is completed, RS produces the best parameter combination for the model based on the

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

training data. This optimal hyperparameter configuration is then applied in the modeling stage to build an optimized RF model.

2.5.2. Research Experiments

The main problem in this study is the limited amount of data and the incomplete attributes in the private dataset obtained from RSUD M. Yunus, Bengkulu City. Therefore, this research aims to improve the quality and performance of the CVD prediction model under limited data availability by incorporating data from another source (public dataset) and by analyzing the effect of differences in the number of attributes between the local dataset and the external dataset. To address this problem, three research experiments were conducted.

a. Experiment on Attribute Reduction

The objective of Experiment on Attribute Reductionis to examine the impact of reducing the number of attributes in the external dataset. There are two scenarios in Experiment on Attribute Reduction. Scenario 1 represents a model trained using 15 attributes, while Scenario 2 represents a model trained using 13 attributes. The selection of 13 attributes in Scenario 2 is intended to match the number of attributes available in the local dataset so that the model comparison reflects the actual condition of the private data. The two attributes removed in this Experiment on Attribute Reductionre alcohol consumption and physical activity. The testing workflow in Experiment on Attribute Reductionbegins by dividing the public dataset into two parts: training data (80%) and testing data (20%). The data distribution for Experiment on Attribute Reductionis presented in Table 3.

Table 3. Data Split for Experiment on Attribute Reduction

Experiment	Training Data	Testing Data
Scenario 1	48.100 Data (-) 24.141 (+) 23.959	12.025 Data (-) 6.035 (+) 5.990
Scenario 2	48.100 Data (-) 24.141 (+) 23.959	12.025 Data (-) 6.035 (+) 5.990

The testing process begins with training the model on the training data using the Random Forest (RF) algorithm, after which the model is evaluated on the testing data for each scenario. The test results are assessed using evaluation metrics to identify the effect of differences in the number of attributes on model performance. Through the testing workflow in Experiment on Attribute Reduction, a clear understanding of the impact of attribute reduction on the performance of the CVD prediction model is expected to be obtained. Visualisasi Experiment on Attribute Reduction dapat dilihat pada Figure 2.

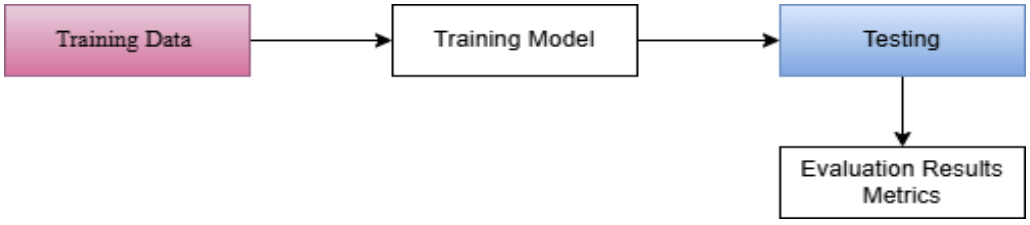

Figure 2. Experiment on Attribute Reduction

b. Experiment Comparing Model Performance with Public and Private Data

The objective of Experiment Comparing Model Performance with Public and Private Data is to compare the performance of prediction models trained using the local dataset and the dataset from

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

another source, followed by testing with evaluation data derived from the local dataset. There are two scenarios in Experiment Comparing Model Performance with Public and Private Data: Scenario 1 uses the external dataset with 13 attributes, while Scenario 2 uses the local dataset. The testing workflow in Experiment Comparing Model Performance with Public and Private Data begins by dividing the dataset into two parts, namely training data and testing data. The data distribution for Experiment Comparing Model Performance with Public and Private Data is presented in Table 4.

Table 4. Data Split for Experiment Comparing Model Performance with Public and Private Data

Experiment	Training Data	Testing Data
Scenario 1	127 Data (-) 64 (+) 63	32 Data (Private Data) (-) 14 (+) 18
Scenario 2	127 Data (-) 64 (-) 64	

This testing process begins by dividing the dataset into five folds, where each fold alternately serves as the testing data (T1–T5) [34]. In Scenario 1, the model is trained using a dataset from another source with 13 attributes and tested on the five folds of private data. In Scenario 2, the model is trained using four folds of the local dataset and tested alternately on the remaining fold of the local dataset that has not been used during training. Each test produces model performance values for predicting CVD. The results from the five tests are then averaged to represent the overall model performance, allowing an objective comparison between models based on the local dataset and those based on a dataset from another source, while reflecting real-world conditions. Visualisasi Experiment Comparing Model Performance with Public and Private Data dapat dilihat pada Figure 3.

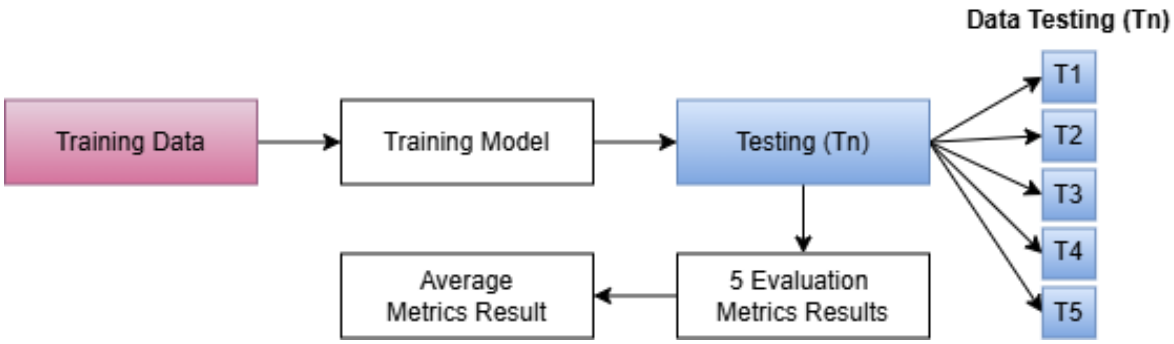

Figure 3. Experiment Comparing Model Performance with Public and Private Data

c. Experiment on Data Integration

The objective of the Experiment on Data Integration was to evaluate the impact of combining limited private data with external public datasets on the quality and performance of the CVD prediction model. In this experiment, the training set was constructed by merging public and private data, while the testing set consisted exclusively of private data. The private dataset was initially divided into 80% training data (127 data) and 20% testing data (32 data). The integration process began by combining 100% of the private training data (127 data) with varying proportions of public data. This study implemented five integration scenarios, from Scenario 1 to Scenario 5, in which the proportion of public data was gradually increased from 50% to 400% the private training data. This design allowed for a systematic evaluation of how different levels of data augmentation influence model performance. The data configuration for Experiment on Data Integration is presented in Table 5.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

Table 5. Data Split for Experiment on Data Integration

Integrated Data Public + Private	Training Data	Testing Data
Scenario 1 50% + 100%	(64 'Public' + 127 'Private') Total = 191 (-) 96 (+) 95	32 Data (Private Data)
Scenario2 100% + 100%	(127 'Public' + 127 'Private') (Total = 254) (-) 127 (+) 127	
Scenario 3 200% + 100%	(254 'Public' + 127 'Private') (Total = 381) (-) 191 (+) 190	
Scenario 4 300% + 100%	(381 'Public' + 127 'Private') (Total = 508) (-) 254 (+) 254	
Scenario 5 400% + 100%	(508 'Public' + 127 'Private') (Total = 635) (-) 318 (+) 317	

In the Experiment on Data Integration, the evaluation process began with the application RFECV on the private dataset to identify the most relevant attributes for prediction. After selecting the optimal features, the private data were processed using a 5-fold cross-validation scheme and then split with a 4:1 ratio (80% training and 20% testing). The private training data were subsequently combined with the public dataset to construct the final training set used for model development. Hyperparameter tuning was then performed using Random Search. The final model generated from the integrated dataset was evaluated using the 20% private test data that had been separated at the beginning of the study to ensure an objective assessment. Model performance was measured using four evaluation metrics accuracy, precision, recall, and F1-score providing a comprehensive yet clear evaluation framework relevant to CVD risk prediction. Visualisasi Experiment on Data Integration dapat dilihat pada Figure 4.

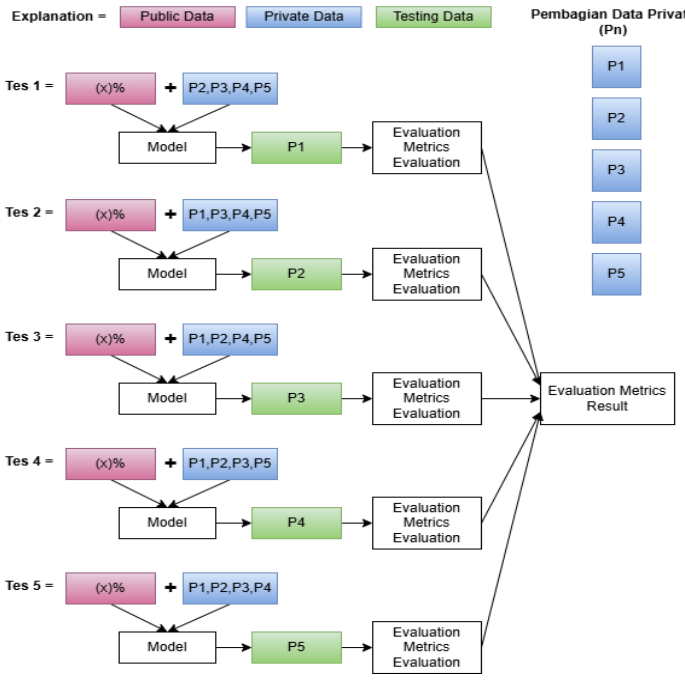

Figure 4. Experiment on Data Integration

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

2.6. Evaluation

The evaluation stage in this study uses four main metrics accuracy, precision, recall, and F1-score to assess the model's ability to correctly identify positive and negative classes, with the evaluation results visualized as comparison graphs for each scenario.

The accuracy metric measures how often the model correctly predicts both positive and negative CVD labels [35].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The precision metric measures how often the model correctly predicts positive CVD cases according to the true labels [35].

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The recall metric measures how effectively the model detects all positive disease cases among the total actual positive cases [35].

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1-score represents the harmonic mean of precision and recall, providing a better indication of the balance between these two metrics.

$$\text{F1} - \text{Scores} = 2\,x\,\frac{Precision \text{ x } Recall}{Precision + Recall} \quad (4)$$

In the final stage, the SHAP method is applied as a validation approach to explain the contribution of each attribute to the prediction results. SHAP values indicate how each attribute increases or decreases the prediction and help identify the most influential attributes in the model's decision making process, thereby ensuring that the predictions are logical, transparent, and aligned with relevant risk factors.

3. RESULT

3.1. Business Understanding

During the Business Understanding stage, an evaluation is conducted to assess the extent to which the research results address the objectives formulated in the research methodology. The results of all experiments indicate that integrating the local dataset with a dataset from another source improves model performance compared with using only a single data source. In addition, reducing the number of attributes does not always decrease model performance as long as the attributes with high contribution are retained. These findings suggest that a CVD prediction model can be developed more efficiently even with fewer attributes, making it more suitable for implementation in healthcare facilities with limited data recording capabilities.

3.2. Data Understanding

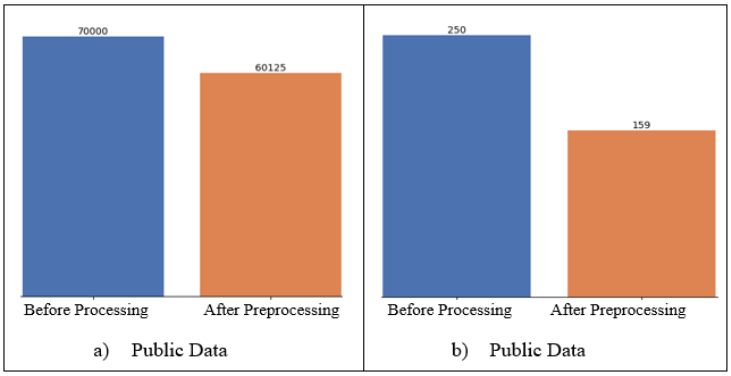

Figure 5. Number of Research Data Records

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

During the Data Understanding stage, an analysis is conducted on the datasets used in this study, including the private dataset and the public dataset. A visualization of the number of records before and after the preprocessing stage is presented in Figure 5.

3.3. Data Preparation

During the data preparation stage, data cleaning is performed to ensure that the dataset is ready for use in the modeling process, allowing the model to learn from the data optimally. Examples of preprocessed data ready for use in the modeling process are shown in Table 6, 7, 8, and 9 below.

Table 6. Sample of Public Data (15 Attribute)

age	gender	height	weight	ap_hi	ap_lo	choles	gluc	smoke	alco	active	bmi	pulse	ap_hi_ ap_lo_ ratio	bp_ category
50	1	168	62	110	80	1	1	0	0	1	21	30	1.375000	0
55	0	156	85	140	90	3	1	0	0	1	34	50	1.555556	1
51	0	165	64	130	70	3	1	0	0	0	23	60	1.857143	0
48	1	169	82	150	10 0	1	1	0	0	1	28	50	1.500000	1
47	0	156	56	100	60	1	1	0	0	0	23	40	1.666667	0

Table 7. Sample of Public Data (13 Attribute)

age	gender	height	weight	ap_hi	ap_lo	choles	gluc	smoke	bmi	pulse	ap_hi_ ap_lo_ratio	bp_ category
50	1	168	62	110	80	1	1	0	21	30	1.375000	0
55	0	156	85	140	90	3	1	0	34	50	1.555556	1
51	0	165	64	130	70	3	1	0	23	60	1.857143	0
48	1	169	82	150	100	1	1	0	28	50	1.500000	1
47	0	156	56	100	60	1	1	0	23	40	1.666667	0

Table 8. Sample of Private Data

age	gender	height	weight	ap_hi	ap_lo	choles	gluc	smoke	bmi	pulse	ap_hi_ ap_lo_ratio	bp_ category
44	20	160	52	140	85	2	1	0	20	55	1.647059	1
65	20	168	59	155	90	2	2	1	20	65	1.722222	1
69	21	170	62	161	95	1	2	0	21	66	1.694737	1
69	22	169	65	150	95	1	3	0	22	55	1.578947	1
54	21	169	60	150	85	1	1	1	21	65	1.764706	1

Table 9. Sample of Integrated Data for Experiment on Data Integration

age	gender	height	weight	ap_hi	ap_lo	choles	gluc	smoke	bmi	pulse	ap_hi_ ap_lo_ratio	bp_ category
50	1	168	62	110	80	1	1	0	21	30	1.375000	0
55	0	156	85	140	90	3	1	0	34	50	1.555556	1
51	0	165	64	130	70	3	1	0	23	60	1.857143	0
48	1	169	82	150	100	1	1	0	28	50	1.500000	1
47	0	156	56	100	60	1	1	0	23	40	1.666667	0

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

3.4. Modelling and Evaluation

The modeling stage is conducted using the Random Forest algorithm with hyperparameter optimization through Random Search (RS), feature selection using Recursive Feature Elimination Cross Validation (RFECV) to identify the most relevant attributes, and model interpretation using SHAP to explain the contribution of each attribute in the prediction process.

The research results are obtained from the implementation of three experiments evaluated using four main metrics: accuracy, precision, recall, and F1-score. This analysis aims to examine the effects of attribute reduction and the integration of the local dataset with a dataset from another source. Overall results from the three research experiments are presented in Table 10.

Table 10. Results of the Research Experiments

No	Experiment	Data	Testing Data	Accuracy	Precision	Recall	F1-Score
1.	Experiment on Attribute Reduction	Scenario 1	Public	72.09	73.63	68.51	70.98
		Scenario 2	Public	71.88	73.81	67.48	70.50
2.	Experiment Comparing Model Performance with Public and Private Data	Scenario 1	Private	48.41	53.85	39.02	45.00
		Scenario 2	Private	70.40	74.24	69.80	71.70
3.	Experiment on Data Integration	Scenario 1	Private	71.67	74.04	74.51	73.85
		Scenario 2	Private	71.65	72.22	79.15	75.26
		Scenario 3	**Private**	**73.57**	**73.66**	**81.44**	**77.06**
		Scenario 4	Private	72.30	72.79	79.15	75.73
		Scenario 5	Private	71.03	71.85	77.97	74.59

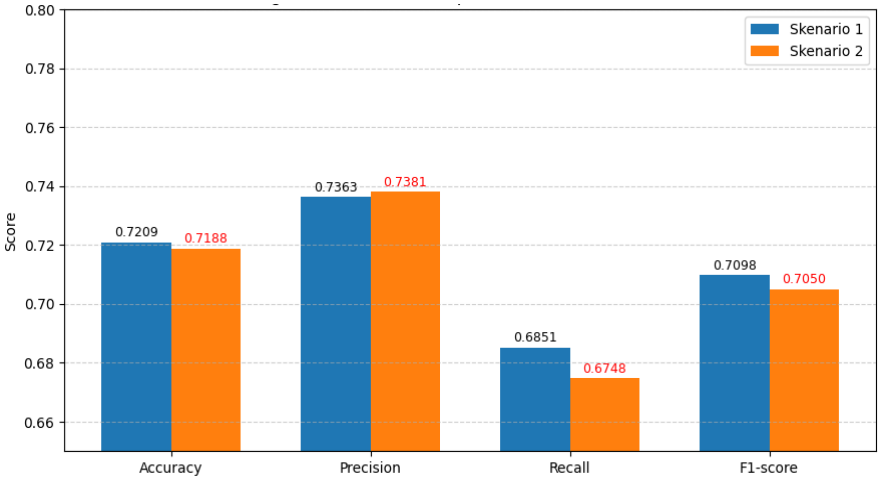

Figure 6. Comparison of Experiment on Attribute Reduction Results

In Experiment on Attribute Reduction, the results showed that the performance difference between Scenario 1 and Scenario 2 was relatively minimal. This finding indicates that the removal of two attributes (alcohol and physical activity) did not have a substantial impact on the performance of the CVD prediction model. In a medical context, a decrease in the F1-score is relevant; however, a reduction of less than 0.5 suggests that the model remains balanced between precision and recall despite the reduction in the number of features. Therefore, the model's ability to detect positive CVD cases did not decline significantly after removing the two attributes. These results suggest that the feature set used

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

in Scenario 2 is sufficiently representative to develop an effective and efficient CVD prediction model. The visualization of Experiment on Attribute Reductionis presented in Figure 6.

After obtaining the results from Experiment on Attribute Reduction, an analysis is conducted to interpret the attributes that most strongly influence the output of the CVD risk prediction model using the SHAP method. The SHAP visualization for Scenario 1 is presented in Figure 7.

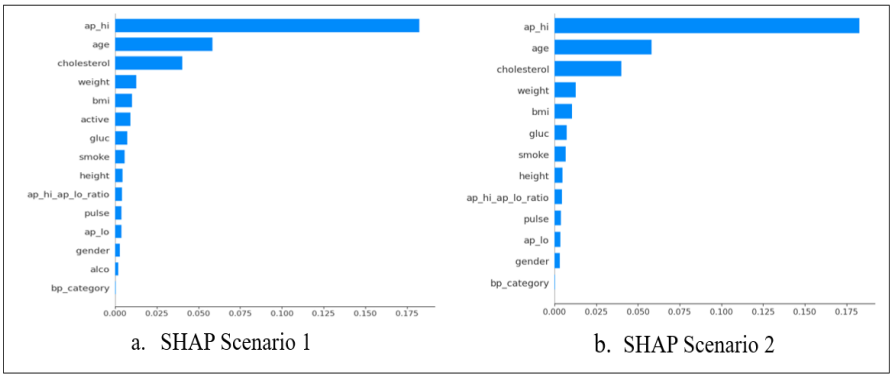

Figure 7. SHAP Results of Experiment on Attribute Reduction

Based on the SHAP results from the two scenarios in Experiment on Attribute Reductionshown in Figure 6, the findings are almost identical: systolic blood pressure (ap_hi) remains the most influential attribute, followed by age and cholesterol. Meanwhile, the other attributes contribute far less than these three variables. This indicates that the removal of two attributes has only a very small effect and does not significantly influence model accuracy or overall performance. Therefore, the SHAP analysis further confirms that removing the active and alco attributes does not meaningfully affect the model's prediction results.

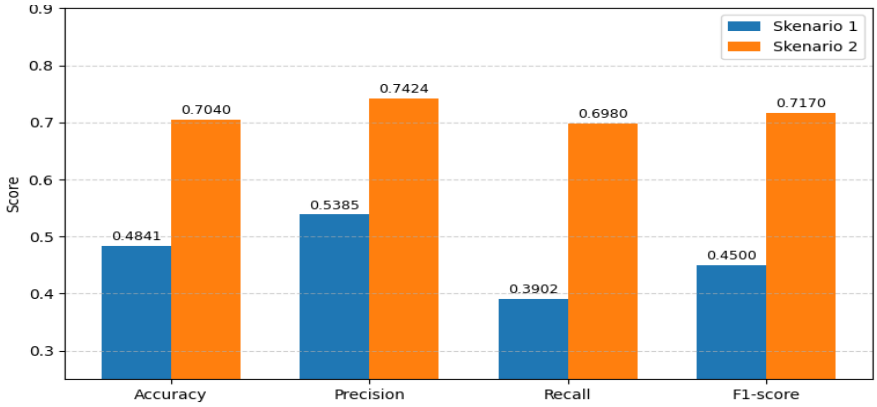

Figure 8. Experiment Comparing Model Performance with Public and Private Data Results

In Experiment Comparing Model Performance with Public and Private Data , the results indicate that Scenario 2 achieved better performance than Scenario 1 when both models were evaluated using private data. The relatively low evaluation scores obtained from the public dataset suggest that it does not adequately represent the characteristics of CVD patients in Indonesia, limiting the model's ability to effectively learn CVD patterns. This limitation may increase the risk of false negatives (patients with the disease predicted as healthy) if not properly addressed. In contrast, the model trained using private data demonstrated a substantially higher recall, indicating a stronger ability to detect CVD cases. Therefore, the performance differences between the two models confirm that the private dataset is more representative of CVD cases in the Indonesian context, as it was collected directly from RSUD M. Yunus

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

Bengkulu. A comparison of the Experiment Comparing Model Performance with Public and Private Data results is presented in Figure 8.

In Experiment on Data Integration, the results show a consistent increase in recall and F1-score from Scenarios 1 to 3, with relatively stable performance. This finding indicates that the gradual addition of public data improves the model's ability to recognize CVD risk patterns when evaluated on private testing data. This improvement is also reflected in the relatively stable and progressively increasing recall and F1-score values, suggesting a good balance between case detection capability and classification accuracy. However, in Scenarios 4 and 5, the evaluation metrics consistently decline, suggesting that the model has exceeded the optimal data integration ratio. Based on the five scenarios conducted in Experiment on Data Integration, Scenario 3 is determined to be the best model because it achieves the highest values across all evaluation metrics. Therefore, the testing process is concluded at Scenario 5. A comparison of the Experiment on Data Integration results is presented in Figure 9.

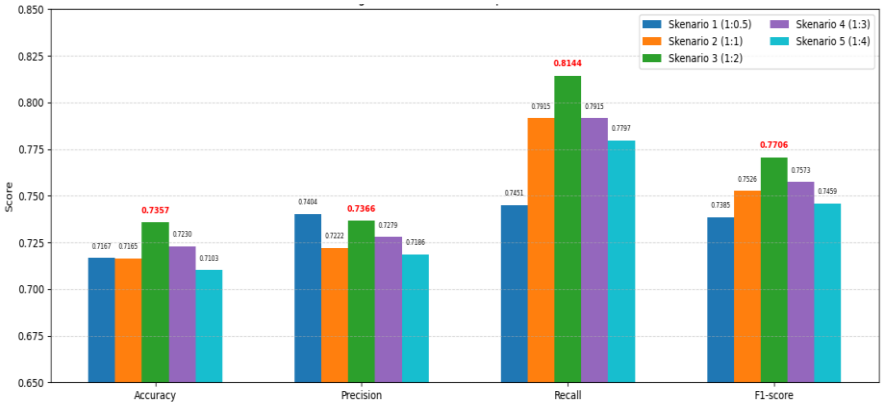

Figure 9. Comparison of Experiment on Data Integration Results

4. DISCUSSIONS

Based on the three experiments conducted, the performance of the CVD prediction model is influenced by the type, quality, and integration strategy of the data. Reducing the number of attributes from 15 to 13 does not produce a significant impact on performance, indicating that the use of 13 attributes is sufficiently efficient without decreasing accuracy. The integration of the local dataset with a dataset from another source is proven to improve model performance when the integration ratio is appropriate.

Based on the SHAP analysis, systolic blood pressure and age were identified as the most influential factors in predicting CVD. From a medical perspective, this finding is consistent with established knowledge, as high blood pressure increases the risk of heart attack and stroke, while advancing age is associated with reduced vascular function. These results indicate that the model successfully captures patterns aligned with clinical evidence, thereby enhancing the reliability of its predictions. From an information systems perspective, the use of SHAP improves model transparency and interpretability, making the results easier for healthcare professionals to understand. This approach has strong potential to support the development of clinical decision support systems in Indonesia. Overall, multisource clinical data integration combined with the RF algorithm, RFECV based feature selection, and SHAP analysis produces a CVD prediction model that is reliable, balanced, and easy to interpret, with the potential to support clinical decision systems for early detection, particularly under conditions of limited local datasets. Different from previous research [7], [17], [20], [24] which only used public datasets to compare ML algorithms, this research integrates local medical record data from RSUD M. Yunus, Bengkulu City, with a public dataset and analyzes the effect of the number of attributes through three experiments.

Therefore, this study contributes by proposing a data integration strategy specifically designed for small-scale clinical datasets, which is highly relevant for many hospitals in Indonesia facing data limitations. This approach is expected to address data limitations and support preventive actions for CVD. Future research is recommended to combine other ML algorithms to further optimize model performance and to utilize larger and more diverse local datasets in order to improve accuracy, stability, and practical implementation in clinical contexts.

5. CONCLUSION

Based on the three experiments conducted, it can be concluded that the performance of the cardiovascular disease (CVD) prediction model is influenced by data type, data quality, and data integration strategy. The integration of the local dataset with a dataset from another source is able to significantly improve model performance when performed with an appropriate ratio. SHAP analysis indicates that blood pressure and age are the most influential factors in predicting CVD risk. The application of RFECV does not eliminate any attributes, indicating that all attributes provide relevant contributions to CVD prediction and that the model operates optimally. Overall, this study demonstrates that multisource clinical data integration with an appropriate ratio, combined with the Random Forest algorithm and SHAP analysis, can produce a CVD prediction model that is reliable, balanced, and easy to interpret.

In addition, this study provides an important contribution to the field of Information Systems, particularly in addressing the challenges of limited clinical datasets in health prediction systems. The proposed data integration strategy offers a practical solution for developing disease detection systems under data constraints and provides policy insights for hospitals and healthcare centers regarding the most influential attributes in detecting cardiovascular disease. By demonstrating that the appropriate integration of local and public data can improve model performance without reducing interpretability, this research supports the development of more robust, transparent, and applicable clinical decision support systems, especially in healthcare environments with limited data availability.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper. This research was conducted independently without any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to RSUD M. Yunus for providing access to the research data. They also thank all individuals who contributed directly or indirectly to the research process and manuscript preparation.

REFERENCES

[1] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular Disease Detection using Ensemble Learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, Aug. 2022, doi: 10.1155/2022/5267498.

[2] L. Ciumărnean *et al.*, "Cardiovascular risk factors and physical activity for the prevention of cardiovascular diseases in the elderly," *Int. J. Environ. Res. Public Health*, vol. 19, no. 1, pp. 207–223, Jan. 2022, doi: 10.3390/ijerph19010207.

[3] E. Dritsas, S. Alexiou, and K. Moustakas, "Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques," in *International Conference on Information and*

Vol. 7, No. 1, February 2026, Page. 691-705
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.1.5744

Communication Technologies for Ageing Well and e-Health, ICT4AWE - Proceedings, Science and Technology Publications, Lda, 2022, pp. 315–321. doi: 10.5220/0011088300003188.

[4] L. A. Kaminsky, C. German, M. Imboden, C. Ozemek, J. E. Peterman, and P. H. Brubaker, "The importance of healthy lifestyle behaviors in the prevention of cardiovascular disease," *Prog. Cardiovasc. Dis.*, vol. 70, pp. 8–15, Jan. 2022, doi: 10.1016/j.pcad.2021.12.001.

[5] Staff of WHO, "Cardiovascular diseases (CVDs)," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[6] Badan Kebijakan Pembangunan Kesehatan (BKPK) Kemenkes, "Survei Kesehatan Indonesia (SKI) 2023," 2023.

[7] D. Fania, I. Waspada, and H. A. Wibawa, "Addressing Data Limitations in Cardiovascular Disease Prediction: Integration of Public Databases and Clinical Records," in *International Conference on Informatics and Computational Sciences (ICICoS)*, Institute of Electrical and Electronics Engineers (IEEE), Jan. 2026, pp. 293–298. doi: 10.1109/icicos68590.2025.11329869.

[8] S. Dalal *et al.*, "Application of Machine Learning for Cardiovascular Disease Risk Prediction," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/9418666.

[9] A. H. Elmi, A. Abdullahi, and M. A. Barre, "A machine learning approach to cardiovascular disease prediction with advanced feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 1030–1041, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1030-1041.

[10] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data," *Med. Eng. Phys.*, vol. 105, Jul. 2022, doi: 10.1016/j.medengphy.2022.103825.

[11] S. D. Reddy, S. Lohitha, and F. Shaik, "Machine Learning based Mobile App for Heart Disease Prediction," in *International Conference on Innovative Data Communication Technologies and Application, ICIDCA 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 464–470. doi: 10.1109/ICIDCA56705.2023.10099714.

[12] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2022.100130.

[13] H. N. Huang *et al.*, "Employing feature engineering strategies to improve the performance of machine learning algorithms on echocardiogram dataset," *Digit. Health*, vol. 9, Jan. 2023, doi: 10.1177/20552076231207589.

[14] C. Dadiyala, A. A. Saxena, K. A. Kale, K. A. Bhattad, N. T. S. Sheikh, and Priyanshi, "Progressive Heart Disease Prediction Model Using Machine Learning: A Comprehensive Staging Approach," in *International Conference on Smart Systems for Applications in Electrical Sciences, ICSSES 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICSSES62373.2024.10561373.

[15] G. Kumar Sahoo, K. Kanike, S. K. Das, and P. Singh, "Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, IEEE Computer Society, Nov. 2022. doi: 10.1109/MLSP55214.2022.9943373.

[16] N. Biswas *et al.*, "Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques," *Biomed Res. Int.*, vol. 2023, 2023, doi: 10.1155/2023/6864343.

[17] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Inform. Med. Unlocked*, vol. 41, Jan. 2023, doi: 10.1016/j.imu.2023.101316.

[18] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.

[19] Y. Yang and H. Wang, "Random Forest-Based Machine Failure Prediction: A Performance Comparison," *Applied Sciences (Switzerland)*, vol. 15, no. 16, Aug. 2025, doi: 10.3390/app15168841.

[20] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Mar. 2021. doi: 10.1088/1742-6596/1817/1/012009.
[21] V. Pandey, U. K. Lilhore, and R. Walia, "A systematic review on cardiovascular disease detection and classification," *Biomed. Signal Process. Control*, vol. 102, Apr. 2025, doi: 10.1016/j.bspc.2024.107329.
[22] K. M. Zobair *et al.*, "Systematic review of Internet of medical things for cardiovascular disease prevention among Australian first nations," Nov. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2023.e22420.
[23] M. D. Christina Magnussen, Ph. D. , Francisco M. Ojeda, M. B. , B. S. , Ph. D. Darryl P. Leong, M. D. Jesus Alegre-Diaz, M. D. , Ph. D. , Philippe Amouyel, and etc, "Global Effect of Modifiable Risk Factors on Cardiovascular Disease and Mortality," *New England Journal of Medicine*, vol. 389, no. 14, pp. 1273–1285, Oct. 2023, doi: 10.1056/NEJMoa2206916.
[24] M. Balakrishnan, A. B. Arockia Christopher, P. Ramprakash, and A. Logeswari, "Prediction of Cardiovascular Disease using Machine Learning," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1767/1/012013.
[25] S. DEMİR and E. K. ŞAHİN, "Assessment of Feature Selection for Liquefaction Prediction Based on Recursive Feature Elimination," *European Journal of Science and Technology*, Sep. 2021, doi: 10.31590/ejosat.998033.
[26] C. Y. Freytes *et al.*, "Recursive Feature Elimination with Cross Validation for Alzheimer's Disease Classification using Cognitive Exam Scores," in *1st International Conference of Intelligent Methods, Systems and Applications, IMSA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 327–332. doi: 10.1109/IMSA58542.2023.10217660.
[27] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," *Journal of Sensor and Actuator Networks*, vol. 12, no. 5, Oct. 2023, doi: 10.3390/jsan12050067.
[28] E. Miranda, S. Adiarto, F. M. Bhatti, A. Y. Zakiyyah, M. Aryuni, and C. Bernando, "Understanding Arteriosclerotic Heart Disease Patients Using Electronic Health Records: A Machine Learning and Shapley Additive exPlanations Approach," *Healthc. Inform. Res.*, vol. 29, no. 3, pp. 228–238, Jul. 2023, doi: 10.4258/hir.2023.29.3.228.
[29] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Dec. 01, 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1186/s40708-024-00222-1.
[30] X. Tusongtuoheti, Y. Shu, G. Huang, and Y. Mao, "Predicting the risk of subclinical atherosclerosis based on interpretable machine models in a Chinese T2DM population," *Front. Endocrinol. (Lausanne).*, vol. 15, 2024, doi: 10.3389/fendo.2024.1332982.
[31] M. Ibrahim, "Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective," *Dhaka University Journal of Applied Science and Engineering*, vol. 7, no. 1, pp. 66–71, Feb. 2023, doi: 10.3329/dujase.v7i1.62888.
[32] Svetlana Ulianova, "Kaggle Dataset, 'Cardiovascular Disease dataset'.," https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.
[33] Y. Zhang and Z. Wang, "Feature Engineering and Model Optimization Based Classification Method for Network Intrusion Detection," Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app13169363.
[34] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Handling Class Imbalance in Health Datasets: A Comparative Study of SMOTE and SMOTEENN with TabNet," in *2025 8th International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang: Institute of Electrical and Electronics Engineers (IEEE), Jan. 2026, pp. 305–310. doi: 10.1109/icicos68590.2025.11329876.
[35] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Pneumothorax Detection System in Thoracic Radiography Images Using CNN Method," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 981–990, Jan. 2025, doi: 10.15294/sji.v11i4.16635.