ILKOM Jurnal Ilmiah Vol. 16, No. 1, April 2024, pp. 68-75
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716



# Evaluation of Machine Learning Models for Predicting Cardiovascular Disease Based on Framingham Heart Study Data

**Ruddy J Suhatril [a,1,*]; Rama Dian Syah [a,2]; Matrissya Hermita [a,3] ; Bhakti Gunawan [a,4] ; Widya Silfianti [a,5]**

[a] *Universitas Gunadarma, Jl. Margonda Raya 100, Depok, Indonesia*
[1] *ruddyjs@staff.gunadarma.ac.id;* [2] *rama_ds@staff.gunadarma.ac.id;* [3] *matrissya@staff.gunadarma.ac.id;* [4] *bhakti@staff.gunadarma.ac.id;* [5] *wsilfi@staff.gunadarma.ac.id*
** Corresponding author*

**Article history:** Received November 20, 2023; Revised February 26, 2024; Accepted April 25, 2024; Available online April 26, 2024

**Abstract**

The Framingham Heart Study Community is a research community that produces data related to Cardiovascular Disease (CVD). This research applies technology to predict CVD using machine learning based on data from the Framingham Heart Study. The eight machine learning algorithms are deployed in this study, they are decision tree, naïve bayes, k-nearest neighbors, support vector machine, random forest, logistic regression, neural network, and gradient boosting. This research uses several stages of research such as load dataset, preprocessing data, data modeling, evaluation of various data modelling, and input new data. The best performance was produced by the random forest model with an accuracy value of 0.84, a precision value of 0.84, a recall value of 0.85, an f1-score value of 0.79 and an AUC value of 0.72. The prediction generated by the proposed machine learning model is high risk or low risk of CVD.

**Keywords:** CVD; Framingham Heart Study; Machine Learning.

## Introduction

The leading cause of death for people worldwide is cardiovascular disease (CVD). The contribution of causes of death by cardiovascular disease accounts for more than 70% of deaths in the world [1], [2]. Risk factors for cardiovascular disease can occur due to poor lifestyle such as diet, smoking, excessive sugar consumption, lack of exercise and obesity [3], [4]. CVD such as coronary heart disease can be detected using electrograms and CT scans. These detection equipment facilities are often very expensive for certain communities, so knowledge of CVD is very important for prevention or early determination of CVD.

The Framingham Heart Study community is now conducting research on the particular cardiovascular illness. The society was established in 1948 with the objective of researching cardiovascular illnesses [5]. Datasets from research results by the Framingham Heart Study were shared with the world community to increase public knowledge about CVD and as research material by researchers around the world. Research related to CVD continues to be carried out to produce technology that can detect the risk of CVD that may occur in humans at certain ages in the future.

Drozdz conducted research on the use of machine learning to predict cardiovascular risk using data from 191 individuals diagnosed with metabolic-associated fatty liver disease (MAFLD). Modeling the data uses Logistic Regression. The performance of the machine model used by Drodz has an accuracy value of 0.85 [6]. His research produced a machine learning model that can identify MAFLD patients who have CVD prevalence. Heart disease prediction using machine learning is also carried out by Apurv. The research employs K-Nearest Neighbor (KNN) and Random Forest machine learning algorithms. His research yielded a KNN accuracy score of 0.86 and a Random Forest accuracy score of 0.81. His machine learning model has the capability to accurately predict individuals who are either susceptible or not susceptible to heart disease [1]. Machine learning technology is also implemented in the field of surgery. Tanwani used the motion2vec algorithm to learn suture motion from surgical videos in the JIGSAWS dataset. The accuracy produced by the motion2vec algorithm has an average accuracy of 85.5% on suture segmentation in surgical operations. The motion2vec algorithm is used in the da Vinci robotic arm for surgical operations [7].

http://dx.doi.org/10.33096/ilkom.v16i1.1952.68-75

The objective of this study is to assess the optimal machine learning model for predicting cardiovascular illness by utilizing data from the Framingham Heart Study. The utilized machine learning model is a supervised model that incorporates multiple pre-existing algorithms. The machine learning model with the highest performance will undergo testing by inputting new data and generating predictions for potential occurrences of CVD.

## Method

The objective of this study is to assess the efficacy of machine learning algorithms in predicting CVD. The dataset utilized is the Framingham heart research data, which was accessible on Kaggle as of August 2023. The employed techniques include naïve bayes, logistic regression, KNN, Neural Network (NN), Support Vector Machine (SVM), decision tree, random forest, and gradient boosting. The methodological procedures employed include dataset loading, data pre-processing, data modelling, and evaluation of several data modeling techniques. The choice of approach can significantly impact the performance of the most effective algorithm for the CVD prediction model depicted in **Figure 1**.

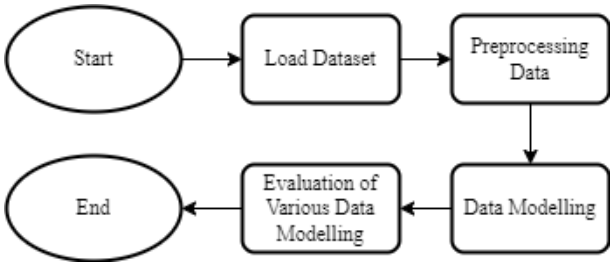


**Figure 1.** Flow Diagram of Research Method

### *A. Load Dataset*

Dataset loading refers to the procedure of extracting a dataset from accessible data sources into software, with the purpose of utilizing the data for data science analysis [8]. The dataset utilized in this work is the Framingham heart study analysis dataset, comprising 4240 patient records with several attributes, as outlined in **Table 1**. The dataset includes the various features.

**Table 1.** Dataset Atribut

| Feature | Variable | Min and Max Value |
|---|---|---|
| Gender | gender | 1: Male and 0: Female |
| Age | age | Min: 32 and max: 70 |
| Current smoker | currentSmoker | 1: yes and 0: no |
| Cigarettes per day | cigsPerDay | Min: 0 and Max: 70 |
| Use of blood pressure medicine | BPMeds | 1: yes and 0: no |
| Stroke | prevalentStroke | 1: yes and 0: no |
| Hypertension | prevalentHyp | 1: yes and 0: no |
| Diabetes | diabetes | 1: yes and 0: no |
| Total cholestrol | totChol | Min: 107 and Max: 696 |
| Systolic blood pressure | sysBP | Min: 83 and Max: 295 |
| Diastolic blood pressure | diaBP | Min: 48 and Max: 142 |
| Body mass index | BMI | Min: 15.54 and Max: 56.8 |
| Hearth rate | heartRate | Min: 44 and Max: 143 |
| glucose | glucose | Min: 40 and Max: 394 |
| CVD at 10 years | TenYearCHD | 1: yes and 0: no |
| Current smoker | currentSmoker | 1: yes and 0: no |

**Table 1** shows the features from the 15 datasets. The categorical features are gender, age, current smoker, stroke, hypertension, diabetes, and CVD at 10 years. The target class for classification is the feature CVD at 10 years which is represented as 1 with status risk of CVD at 10 years or 0 with status no risk of CVD at 10 years.

### *B. Preprocessing Data*

Data preprocessing is an important process in data science before the data is used. Raw data is transformed into clean data using data preprocessing. [9]. The purpose of data preprocessing is to check for problematic data, inconsistent

data, and missing data. The data preprocessing technique used is the missing value imputation technique. Deleting data that is empty or has a NaN value is done using the missing value imputation technique [10]. **Table 2** shows the amount of empty data for each variable.

**Table 2.** Missing Value

| Variable | Missing Value | Mssing Value After Preprocessing |
|---|---|---|
| gender | 0 | 0 |
| age | 0 | 0 |
| currentSmoker | 0 | 0 |
| cigsPerDay | 29 | 0 |
| BPMeds | 53 | 0 |
| prevalentStroke | 0 | 0 |
| prevalentHyp | 0 | 0 |
| diabetes | 0 | 0 |
| totChol | 50 | 0 |
| sysBP | 0 | 0 |
| diaBP | 0 | 0 |
| BMI | 19 | 0 |
| heartRate | 1 | 0 |
| glucose | 388 | 0 |
| TenYearCHD | 0 | 0 |

### *C. Modelling Data*

Data modeling in machine learning aims to represent data mathematically to carry out training and testing on the dataset used [11]. The data modeling process can identify patterns, trends, or relationships in data that can be used to make predictions [12]. The dataset is used to provide training and testing data. Testing data is used to determine the model performance after it has been trained using training data. The proportion of applied training to testing data is 75% for training and 25% for testing. Several machine learning algorithms are used in this study as in **Figure 2**.

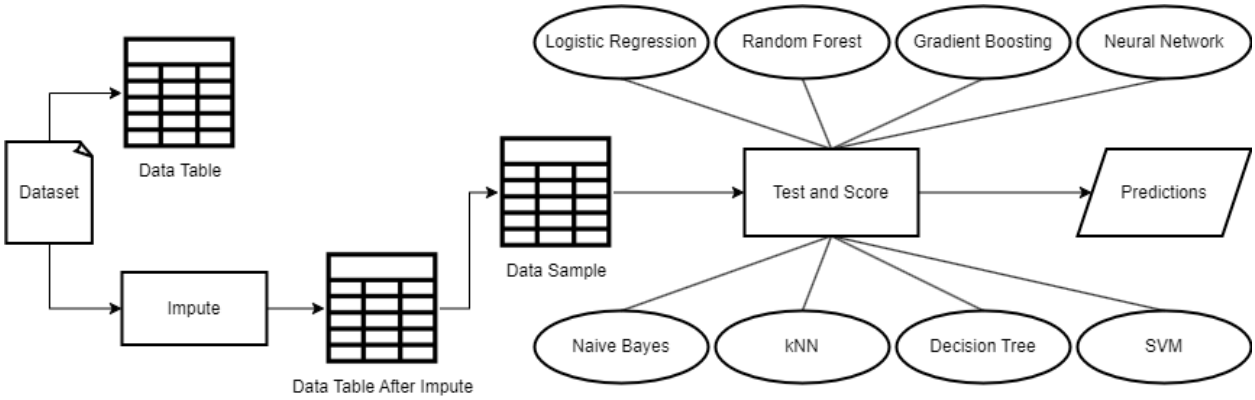


**Figure 2.** Model Machine Learning

**Figure 2** shows the process of machine learning modeling with many algorithms. The dataset is provided as input through the file widget. The impute widget is used to perform data preparation and address missing values in the dataset. The data sampler widget is responsible for partitioning the training data and testing data. The machine learning methods are represented by the logistic regression, naïve bayes, tree, KNN, SVM, random forest, NN, and gradient boosting widgets. The performance of each method was evaluated using the test score widget. The prediction widget is utilized to do testing data predictions.

#### *1) Random Forest*

The random forest technique, a form of supervised machine learning, is valuable for both regression and classification tasks. The Random Forest technique utilizes many decision trees and aggregates the results of these trees to get predictions that are more precise and consistent [13]. The random forest method offers numerous benefits, including a consistently high level of accuracy, exceptional efficiency in handling large datasets, comprehensive calculation of all classification features, and the ability to prevent overfitting. Random forest has several parameter [14]. The n_estimator parameter is used to determine the number of decision trees

modelled. The max_feature parameter is used to determine the number of features. The min_sample_split parameter is used to determine the minimum number of samples that will be required to split an internal node. The min_sample_leaf parameter is used to determine the minimum number of samples at the leaf node. The max_depth parameter is used to determine the depth of each tree and prevent overfitting [15].

*2) Decision Tree*

Decision trees are frequently utilized in a variety of environments, including pattern recognition, image processing, and machine learning [16]. Decision trees consist of trees where each node represents a question or test against certain data attributes. The results of these questions will go to the appropriate child node and will carry out the next question or test. The process will proceed until it reaches a leaf node in order to generate a definitive decision or forecast. The decision tree model will generate an entropy value. The entropy value will fluctuate when the training data is partitioned into smaller groups using tree nodes. The information gain value is the difference in entropy value that occurs as a result.

*3) Naïve Bayes*

Thomas Bayes is the scientist who is acknowledged for the invention of the naïve Bayes algorithm. The Naïve Bayes algorithm utilizes probability and statistical principles to predict the probability of future events based on past events [17]. This approach has the advantage of using a small quantity of training data to determine the parameters needed for the classification process [18].

*4) K-Nearest Neighbors*

The classification process of the KNN algorithm is directly affected by the training dataset [19]. The KNN technique selects the K nearest points to the testing query by utilizing a training model that closely resembles the query, as a preliminary step for classification. The ultimate classification will be determined based on the majority of votes cast [20].

*5) Support Vector Machine*

The discovery of SVM is credited to an American scientist named Vapnik. The operational mechanism of SVM for classification involves determining the optimal hyperplane or dividing plane that effectively separates two groups of data. The hyperplane is determined by maximizing the margin measurement. The margin denotes the smallest distance between data points that belong to distinct classes and the hyperplane. The data points or vectors that are in closest vicinity to the hyperplane have the capacity to influence the positioning of the hyperplane, and hence, they are designated as support vectors.

*6) Logistic Regression*

The Logistic Regression algorithm is a machine learning technique used for analyzing and modeling data, particularly in classification tasks. Logistic Regression in machine learning models evaluates the connection between categorical target variables with nominal scale levels and categorical predictor variables with interval or ratio scale levels [21]. Logistic regression algorithm can be utilized to establish correlations between variables in a time series model. The logistic regression algorithm can be used to predict the probability of the dependent variable falling into several categories.

*7) Neural Network*

The neural network algorithm is a machine learning method that shares similarities with the biological neural network systems found in humans [22]. A neural network is a mathematical model composed of interconnected layers of artificial neurons. The neural network consists of nodes, including cell bodies, which are capable of communication with other nodes through connections resembling axons and dendrites. This method uses modeling techniques to identify patterns in data and establish the correlation between input and output.

*8) Gradient Boosting*

Gradient Boosting algorithm is included in the supervised machine learning category which is based on decision trees in making predictions. Gradient Boosting uses the combination of several weak models such as decision trees into a stronger model [23]. This algorithm always maximizes an objective function such as the MSE function for regression or the Log-Loss function for classification by evaluating the gradient at each point. This algorithm takes an iterative approach to improve the previous model by adding new models until the resulting model meets certain criteria.

## Results and Discussion

This research was conducted on a computer with Intel(R) Core (TM) i7-4940 @ 3.60GHz Processor and 16 GB RAM. After preprocessing, the 4240 records in the dataset were reduced to 3658 records and 15 features. The dataset consists of 25% testing data and 75% training data.

### *A. Evaluation of various data modeling*

This study employs multiple machine learning methods to predict CVD. The algorithm's performance is evaluated based on metrics such as F1 Score, recall, accuracy, and precision. The evaluation metrics for each model are displayed in **Table 3**. The Random Forest model achieves the greatest accuracy among all machine learning models, with a value of 0.85. Additionally, it exhibits precision, recall, and F1 Score values of 0.84, 0.85, and 0.79, respectively.

**Table 3.** Evaluation Results from Various Models

| Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.77 | 0.77 | 0.77 | 0.77 | 0.58 |
| Naïve Bayes | 0.81 | 0.77 | 0.81 | 0.78 | 0.72 |
| K-Nearest Neighbors | 0.83 | 0.77 | 0.83 | 0.78 | 0.63 |
| Support Vector Machine | 0.84 | 0.70 | 0.84 | 0.76 | 0.58 |
| Random Forest | 0.85 | 0.84 | 0.85 | 0.79 | 0.70 |
| Neural Network | 0.83 | 0.79 | 0.84 | 0.78 | 0.69 |
| Logistic Regression | 0.84 | 0.78 | 0.84 | 0.77 | 0.71 |
| Gradient Boosting | 0.83 | 0.77 | 0.84 | 0.78 | 0.70 |

### *B. ROC Curve Evaluation*

Receiver Operating Characteristic (ROC) curve evaluation was carried out to determine the graphical relationship between the true positive rate and the false positive rate in each machine learning model. ROC evaluation aims to determine the performance of the classification model in distinguishing between positive and negative classes. AUC is a measuring tool used to determine the diagnostic accuracy value that represents the subject correctly [24]. It can be seen in Figure 3 that the highest AUC performance value is found in the Naïve Bayes model, with a value of 0.72.

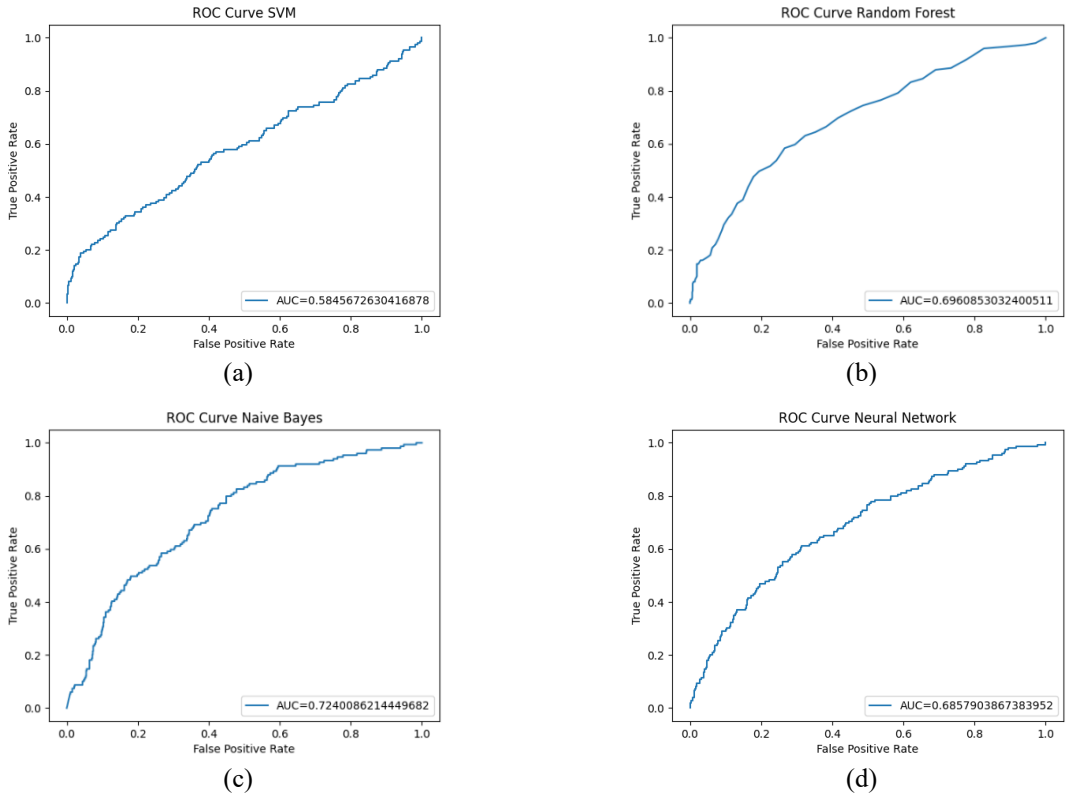


**Figure 3.** ROC Curve (a) SVM, (b) Random Forest, (c) Naïve Bayes, (d) Neural Network.

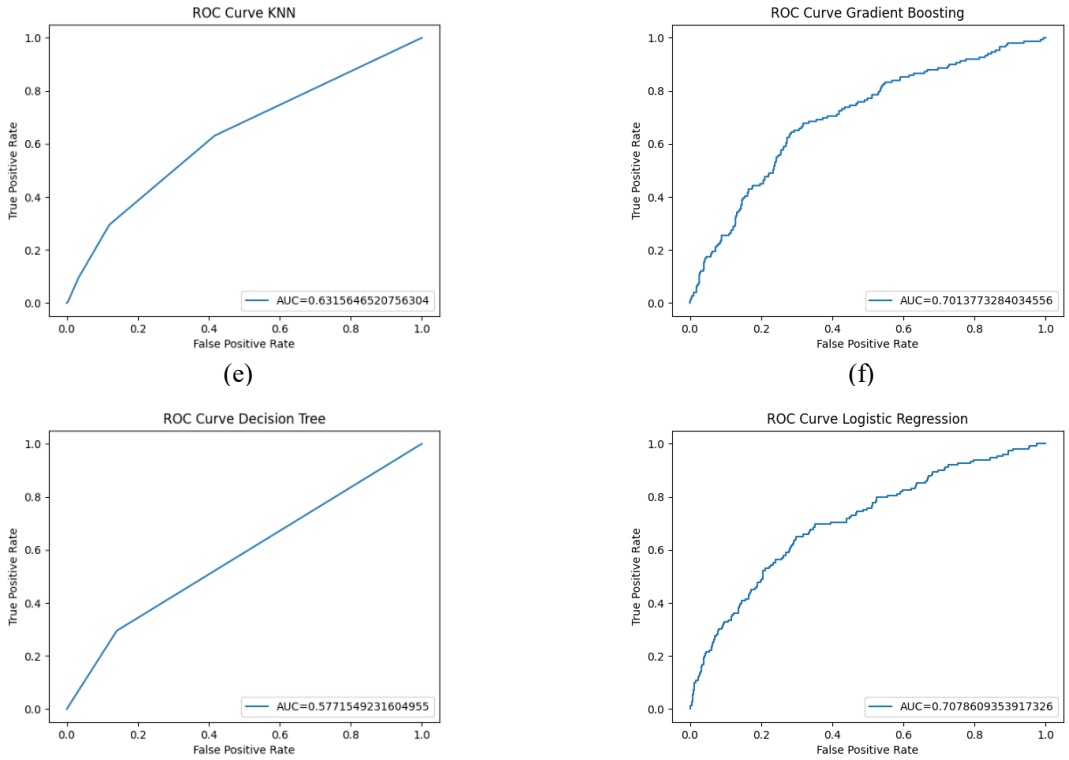


**Figure 3.** ROC Curve (e) KNN, (f) Gradient Boosting, (g) Decision Tree, and (h) Logistic Regression

### *C. Testing New Data Input*

Prediction for new input data were carried out on the Random Forest Model which has highest accuracy value. It can be seen in **Figure 4** that the input data is different and produces results with different statuses depending on the condition of the input data. The resulting prediction results can be 'Low risk of CVD in Ten Years' or 'High risk of CVD in Ten Years'.

```
Gender (0 = Women | 1 = Men): 0
Age: 55
Smoke (1=Yes | 0=No): 0
Cigs Per Day: 0
Consuming Blood Pressure Medicine (1=Yes | 0=No): 0
Stroke (1=Yes | 0=No): 1
Hypertensi (1=Yes | 0=No): 0
Diabetes (1=Yes | 0=No): 0
Cholestrol: 189
Systolic Blood Pressure: 120
Diastolic Blood Pressure: 95
BMI: 22.9
Heart Rate: 100
Glucosa: 115
**********
Result: Low risk of CVD in Ten Years
```

```
Gender (0 = Women | 1 = Men): 0
Age: 55
Smoke (1=Yes | 0=No): 1
Cigs Per Day: 20
Consuming Blood Pressure Medicine (1=Yes | 0=No): 1
Stroke (1=Yes | 0=No): 1
Hypertensi (1=Yes | 0=No): 1
Diabetes (1=Yes | 0=No): 1
Cholestrol: 20
Systolic Blood Pressure: 140
Diastolic Blood Pressure: 100
BMI: 25
Heart Rate: 110
Glucosa: 210
**********
Result: High risk of CVD in Ten Years
```

**Figure 4.** Input New Data

## Conclusion

This study uses the Framingham Heart Study dataset along with multiple classification models to predict CVD. The training and testing datasets are separated by a percentage of 75:25. The Random Forest model with accuracy, precision, recall, and f1 scores of 0.85, 0.84, and 0.79, is the best performing machine learning model. The Naïve Bayes model shows the best AUC value performance with an AUC value of 0.72. Testing of new input data is carried out using the model with the highest accuracy, namely Random Forest. The prediction depends on several data such as gender, age, smoker status, number of cigarettes per day, blood pressure medication consumption, history of stroke, history of hypertension, history of diabetes, cholesterol value, diastolic blood pressure value, systolic blood pressure value, BMI value, heart rate, and glucose value. The status of the prediction results is Low Risk of CVD or High Risk of CVD.

## References


[1] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 12046, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012046.

[2] N. R. Panda, K. L. Mahanta, J. K. Pati, R. Bhuyan, and S. S. Satapathy, "The Effectiveness of Machine Learning Systems' Accuracy in Predicting Heart Stroke Using Socio-Demographic and Risk Factors-A Comparative Analysis of Various Models," *Natl. J. Community Med.*, vol. 14, no. 6, pp. 371–378, 2023, doi: 10.55489/njcm.140620233026.

[3] R. Agrawal, M. K. Gourisaria, and S. K. Panda, "Diagnosis of Heart Stroke Using Feature Extraction and Sequential Learning Based Models," in *2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22)*, 2022, pp. 1–7, doi: 10.1109/ICETET-SIP-2254415.2022.9791712.

[4] World Heart Report 2023, "World Heart Report 2023 Confronting the World's Number One Killer," 2023.

[5] F. Ewbank, J. Birks, and D. Bulters, "The association between acetylsalicylic acid and subarachnoid haemorrhage: the Framingham Heart Study.," *Sci. Rep.*, vol. 13, no. 1, p. 6533, Apr. 2023, doi: 10.1038/s41598-023-33570-9.

[6] K. Drożdż *et al.*, "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach," *Cardiovasc. Diabetol.*, vol. 21, no. 1, pp. 1–12, 2022, doi: 10.1186/s12933-022-01672-9.

[7] A. K. Tanwani, P. Sermanet, A. Yan, R. Anand, M. Phielipp, and K. Goldberg, "Motion2Vec: Semi-Supervised Representation Learning from Surgical Videos," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 2174–2181, 2020, doi: 10.1109/ICRA40945.2020.9197324.

[8] M. Cordeiro-Costas, D. Villanueva, P. Eguía-Oller, M. Martínez-Comesaña, and S. Ramos, "Load Forecasting with Machine Learning and Deep Learning Methods," *Appl. Sci.*, vol. 13, no. 13, 2023, doi: 10.3390/app13137933.

[9] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Chapter 5 - Diagnosing of disease using machine learning," in *Machine Learning and the Internet of Medical Things in Healthcare*, Academic Press, 2021, pp. 89–111.

[10] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, 2021, doi: 10.3389/fenrg.2021.652801.

[11] N. A. Mahoto, A. Shaikh, A. Sulaiman, M. S. Al Reshan, A. Rajab, and K. Rajab, "A machine learning based data modeling for medical diagnosis," *Biomed. Signal Process. Control*, vol. 81, p. 104481, 2023, doi: 10.1016/j.bspc.2022.104481.

[12] S. Gocheva-Ilieva, "Statistical Data Modeling and Machine Learning with Applications," *Mathematics*, vol. 9, no. 23, 2021, doi: 10.3390/math9232997.

[13] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[14] A. Baita, I. A. Prasetyo, and N. Cahyono, "Hyperparameter Tuning on Random Forest for Diagnose Covid-19," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, pp. 138–143, 2023, doi: 10.33387/jiko.v6i2.6389.

[15] P. Contreras, J. Orellana-Alvear, P. Muñoz, J. Bendix, and R. Célleri, "Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment," *Atmosphere (Basel).*, vol. 12, no. 2, 2021, doi: 10.3390/atmos12020238.

[16] B. Jijo and A. Mohsin Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, pp. 20–28, Jan. 2021, doi: 10.38094/jastt20165.

[17] A. Nazir, A. Akhyar, M. Ramadhani, and Herlina, "Naive Bayes Method for Classification of Student Interest Based on Website Accessed," *J. Phys. Conf. Ser.*, vol. 1655, no. 1, p. 12104, Oct. 2020, doi: 10.1088/1742-6596/1655/1/012104.

[18] R. Achmad and A. S. Girsang, "Implementation of naive bayes classifier algorithm in classification of civil servants," in *Journal of Physics: Conference Series*, 2020, vol. 1485, no. 1, doi: 10.1088/1742-6596/1485/1/012018.

[19] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *{ACM} Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2021, doi: 10.1145/3459665.

[20] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, no. 1, p. 6256, 2022, doi: 10.1038/s41598-022-10358-x.

[21] T. Ciu and R. S. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (International J. New Media Technol.*, vol. 7, no. 1, pp. 33–38, 2020, doi: 10.31937/ijnmt.v7i1.1340.

[22] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. Peter Campbell, "Introduction to machine learning, neural networks, and deep learning," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 1–12, 2020, doi: 10.1167/tvst.9.2.14.

[23] N. Aziz, E. A. P. Akhir, I. A. Aziz, J. Jaafar, M. H. Hasan, and A. N. C. Abas, "A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems," in *2020 International Conference on Computational Intelligence (ICCI)*, 2020, pp. 11–16, doi: 10.1109/ICCI51257.2020.9247843.

[24] S. Parodi, D. Verda, F. Bagnasco, and M. Muselli, "The clinical meaning of the area under a receiver operating characteristic curve for the evaluation of the performance of disease markers.," *Epidemiol. Health*, vol. 44, p. e2022088, 2022, doi: 10.4178/epih.e2022088.