Comparison of Classification Algorithm in Predicting Stroke Disease

[1]Fenna Kemala Hutabarat, [2]Daniel Ryan Hamonangan Sitompul, [3]Stiven Hamonangan Sinurat,
[4]Andreas Situmorang, [5]Ruben, [6]Dennis Jusuf Ziegel, [7]Evta Indra*
[1,2,3,4,5,6,7] Sistem Informasi, Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia
Jl. Sampul No.3, Sei Putih Bar., Medan Petisah, Kota Medan
E-mail : *evtaindra@unprimdn.ac.id

ABSTRAK- To prevent stroke, we need a way to predict whether someone has had a stroke through medical parameters. With the influence of technology in the medical world, stroke can be predicted using the Data Science method, which starts with Data Acquisition, Data Cleaning, Exploratory Data Analysis, Preprocessing, and the last stage is Model Building. Based on the model that has been made, it is concluded that the algorithm with the best performance, in this case, is XGBoost with a precision value of 0.9, a recall value of 0.95, an f1 value of 0.92, and a ROC-AUC value of 0.978 after receiving five folds of cross-validation. With these results, the model created can be used to make predictions in real-time.

Kata kunci : *Machine Learning, Logistic Regression, Random Forest, XGBoost, Stroke*

1. INTRODUCTION

A stroke is a condition where blood flow to the brain is blocked which can cause cell death [1]. According to data from the World Health Organization (WHO), stroke is the second leading cause of death worldwide, accounting for 11% of total deaths. A stroke is very deadly if it cannot be treated quickly and appropriately [2]. To prevent stroke, we need a way to predict whether someone has had a stroke through medical parameters. With the influence of technology in the medical world, stroke can be predicted using Machine Learning methods based on classification algorithms, so that with predetermined medical parameters, the Machine Learning model that has been created can predict whether a person has a stroke.

Many previous studies that have carried out disease prediction have been carried out, for example, research [3], namely "Prediction of Heart Disease Using Machine Learning," and research [4], namely "Long Short-Term Memory Recurrent Neural Network for Stroke Prediction." Research [3] presents a Neural Network model with Multi-Level Perceptron (MLP) that can predict heart disease. The system's output in research [3] will provide predictive results in the form of "YES/NO." Research [4] presents a Recurrent Neural Network model with Long Short-Term Memory (LSTM) that can predict stroke. The output of the research [4] is a detailed result of the model that has been made, such as the value of Precision, Recall, F1, and Accuracy.

Based on the description of the problem described in the previous paragraph, to be able to perform early detection of stroke, the authors suggest making an analysis of medical parameters that can cause stroke and making Machine Learning models that can predict stroke. Parameters used in predicting stroke include age to smoking status. In this research, several classification algorithms will be made, such as Logistic Regression, Random Forest Classifier, and Extreme Boosting Classifier. The algorithm with the best precision, recall, F1, and ROC-AUC values will be the best algorithm in this case.

2. METHODOLOGY

This research was conducted at the Data Analyst Laboratory of Prima Indonesia University. The notebook used in this research is Google Collaboratory. The workflow of this research can be seen in Figure 1. This research began with data collection from the website of the dataset provider, namely Kaggle; after the data is obtained, the Data Cleaning process is carried out to normalize the data; then the Exploratory Data Analysis (EDA) stage is carried out, where dataset visualization will be carried out; after the EDA is carried out, the preprocessing stage will be carried out to prepare the training model; the last stage is Model Building, where the classification algorithm models such as Logistic Regression, Random Forest Classifier, and Extreme Boosting Classifier.

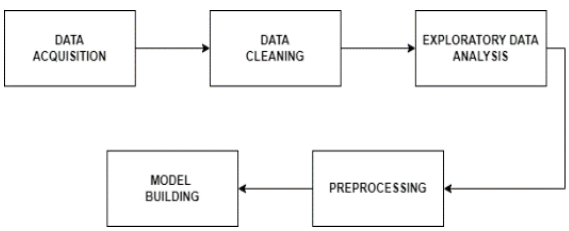

Figure 1. Research Methodology

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_gluco
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92

Figure 2. Dataset Detail

2.1. *Data Acquisition*

The Data Acquisition stage takes data from good sources before the data is processed and used to create a Machine Learning model. This process ends by calling the Data Frame on the Notebook [5], [6]. In this study, the dataset used comes from Kaggle, from Fedesoriano's "Stroke Prediction Dataset" repository [7]. This dataset contains 5110 rows of data and 12 columns which are parameters/features/factors that cause a stroke. Details of the dataset can be seen in Figure 2.

2.2. *Data Cleaning*

The Data Cleaning stage is the process of detecting, repairing, or deleting datasets for analysis preparation (EDA) [8], [9]. The data in question is invalid data that can harm modeling and analysis. In this study, the data cleaning process was carried out to fill in empty data (NULL VALUE) contained in the BMI column in the dataset with an average value (Mean). Details of this stage can be seen in Figure 3.

```
# Mengisi Data Kosong Pada BMI
mean_bmi = stroke.groupby(['gender', 'age_group']).agg({'bmi': 'mean'}).reset_index()
for i in range(len(stroke.index)):
    if pd.isna(stroke.iloc[i, 9]) == True:
        for j in range(len(mean_bmi.index)):
            if mean_bmi.iloc[j, 0] == stroke.iloc[i, 1] and mean_bmi.iloc[j, 1] == stroke.iloc[i, 12]:
                stroke.iloc[i, 9] = mean_bmi.iloc[j, 2]
```

Figure 3. Data Cleaning Process

2.3. *Exploratory Data Analysis (EDA)*

The Exploratory Data Analysis stage is an approach to analyzing the data and summarizing its characteristics of the data. Graphics and visualizations are commonly used in EDA. The primary purpose of EDA is to test the content of non-modeling data to explain hypothesis testing [10], [11]. In this study, the EDA stages include the distribution of stroke by sex to stroke by age and heart disease. Some examples of these stages can be seen in Figure 4.

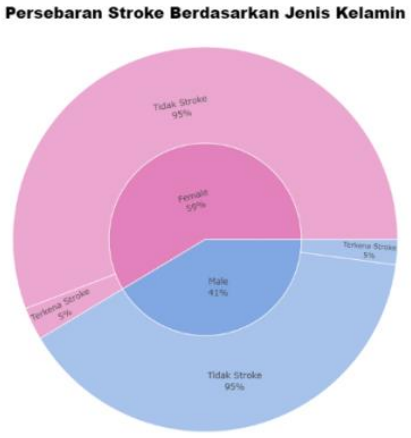

Figure 4. Exploratory Data Analysis Process

2.4. *Preprocessing*

The preprocessing stage prepares relevant data to build and train Machine Learning models. This step also means converting the data into a more readable format for the model. Preprocessing also helps improve data quality [12], [13]. In this study, the Preprocessing stage includes Label Encoding. Details of this stage can be seen in Figure 5.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level
0	1	1.051434	-0.328602	4.185032	1	2	1	2.706375
1	0	0.786070	-0.328602	-0.238947	1	3	0	2.121559
2	1	1.626390	-0.328602	4.185032	1	2	0	-0.005028
3	0	0.255342	-0.328602	-0.238947	1	2	1	1.437358
4	0	1.582163	3.043196	-0.238947	1	3	0	1.501184

Figure 5. Preprocessing

2.5. *Model Building*

This study will use three classification algorithms to predict stroke from medical parameters. The three algorithms used are Logistic Regression, Random Forest Classifier (RFC), and Extreme Gradient Boosting Classifier (XGBoost).

2.5.1. *Logistic Regression*

Logistic Regression is one of the Machine Learning algorithms used to carry out the classification process. This algorithm calculates the probability of a dataset consisting of dependent and independent variables. Since the results used as outputs are opportunities, the dependent variable is limited to values 0 and 1 [14], [15]. In this study, the Logistic Regression algorithm was created using the random_state 22 configurations; the rest is the default configuration.

```
lg = LogisticRegression(random_state = 22)
lg.fit(X_train_balanced, Y_train_balanced)
y_pred = lg.predict(X_test)
y_prob = lg.predict_proba(X_test)[:,1]
```

Figure 6. Logistic Regression Model Making

2.5.2. Random Forest (RF)

Random Forest is a decision tree-based Machine Learning algorithm used to perform regression and classification processes. Random Forest consists of many decision trees that work in groups. Each decision tree in this algorithm produces class predictions, and the class with the most choices becomes the prediction result of the model [16], [17]. This study created Random Forest with random_state 22 and max_depth five configurations.

```
[ ] rf = RandomForestClassifier(random_state = 22, max_depth = 5)
    rf.fit(X_train_balanced, Y_train_balanced)
    y_pred = rf.predict(X_test)
    y_prob = rf.predict_proba(X_test)[:,1]
```

Figure 7. Random Forest Model Making

2.5.3. Extreme Gradient Boosting (XGBoost)

XGBoost is an implementation of the Gradient Boosted Decision Tree. In this algorithm, the decision tree is made in sequential form. Weights have an essential role in XGBoost. Weights are loaded on all independent variables used to make predictions [18], [19]. In this study, the XGBoost algorithm was created using the configuration random_state 22, max_depth 5, objective "binary:logistic," eval_metric "logloss."

```
[ ] xgb = XGBClassifier(random_state = 22, max_depth = 5,
                        objective = 'binary:logistic',
                        eval_metric = 'logloss')
    xgb.fit(X_train_balanced, Y_train_balanced)
    y_pred = xgb.predict(X_test)
    y_prob = xgb.predict_proba(X_test)[:,1]
```

Figure 8. XGBoost Model Making

3. RESULT AND DISCUSSION

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis in this study will be divided into six parts, namely visualization of the distribution of strokes based on gender, visualization of age who are commonly affected by stroke, visualization of BMI levels that are commonly affected by stroke, visualization of glucose levels that are commonly affected by stroke, visualization of stroke comparisons based on hypertension and visualization comparison of stroke by heart disease.

The first part of the distribution of stroke by gender can be seen in Figure 9. From the diagram, it can be concluded that the dataset used in this study has 59% female data and 41% male data. Of the 59% of female data, 95% did not have a stroke, and 5% had a stroke. Of the 41% of male data, 95% did not have a stroke, and 5% had a stroke.

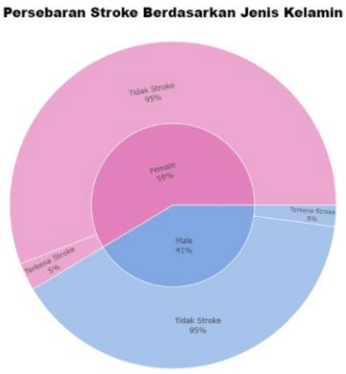

Figure 9. Stroke Based on Gender

In the second part, a visualization of the age commonly affected by stroke can be seen in Figure 10. From the diagram, it can be concluded that ages 0 to 20 years are generally not prone to stroke; the age most susceptible to stroke is 40 to 80. Ages 75-80 80 years is the age most prone to stroke based on the data in the dataset.

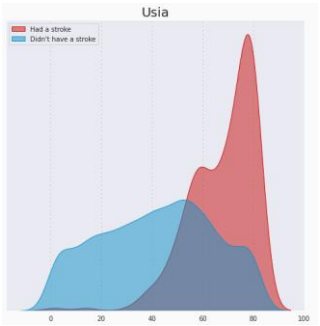

Figure 10. Stroke Based on Age

In the third part, a visualization of the BMI level that generally suffers a stroke can be seen in Figure 11. From the diagram, it can be concluded that someone with a BMI of 30 has a higher chance of stroke than other BMI levels.

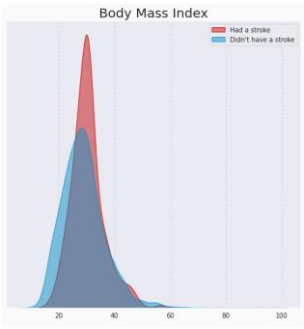

Figure 11. Stroke Based on BMI

In the fourth part, the visualization of glucose levels that are commonly affected by stroke can be seen in Figure 11. From the visualization it can be concluded that glucose levels in the range of 50-100

are more in people who do not have strokes and glucose levels in the range of 200-250 are the most common glucose levels. susceptible to stroke.

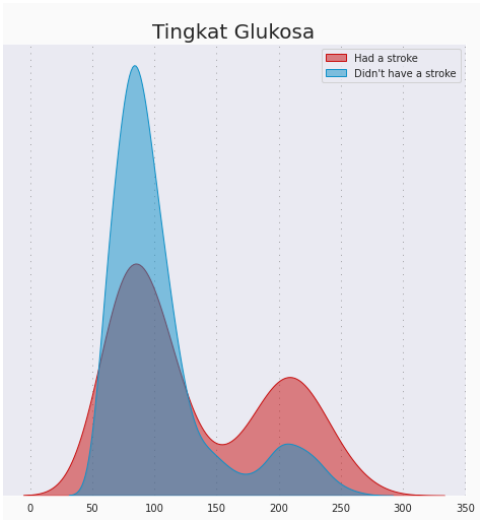

Figure 12. Stroke Based on Glucose Level

In the fifth section, a visualization of the comparison of strokes based on hypertension can be seen in Figure 13. From this visualization, it can be concluded that someone with hypertension problems is more prone to stroke than people who do not have hypertension problems. There are 13.3% of people who have strokes from people with hypertension, and there are 4% of people who have strokes from people who do not have hypertension

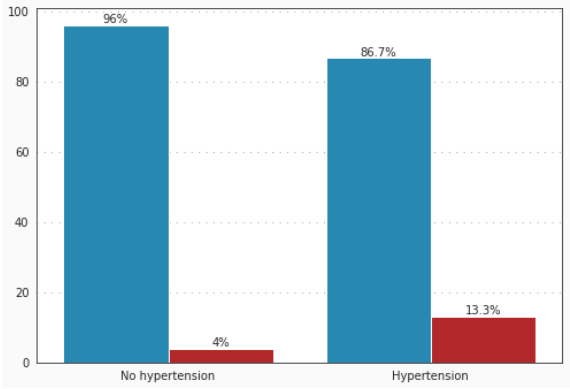

Figure 13. Stroke Based on Hypertension

In the sixth section, a comparison visualization of stroke based on heart disease can be seen in Figure 14. From the visualization, it can be concluded that people with heart disease are more prone to stroke than people without heart disease. There are 17% of people who have had a stroke and also have heart disease, and there are 4.2% of people who have had a stroke but do not have heart disease.

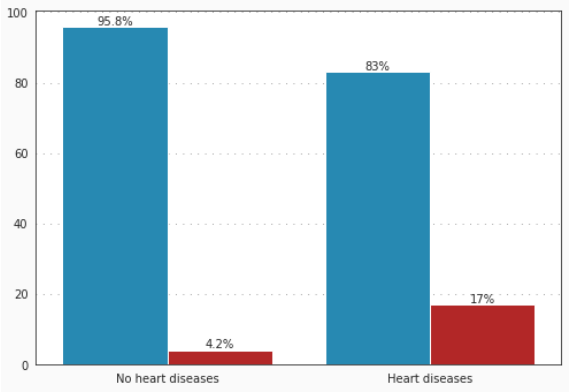

Figure 14. Stroke Berdasarkan Penyakit Jantung

3.2 Feature Importance

Features Importances in this study are determined based on the algorithm used. From the Logistic Regression algorithm, the three most important features are age, gender, and type of residence; from the Random Forest algorithm, the three most important features are age, age group, and glucose level; and for the XGBoost algorithm, the three most important features are age, glucose level group and age group. Details of the results can be seen in Table 1.

Table 1. Feature Importances Detail

Algorithm	Feature Importances	Value
Logistic Regression	*age*	2.134178
	gender	0.782076
	residence_type	0.648289
Random Forest	*age*	0.584485
	age_group	0.086460
	avg_glucose_level	0.056845
XGBoost	*age*	0.250937
	glucose_group	0.081891
	age_group	0.078457

3.3 Model Building Result

3.3.1 Logistic Regression

The results of making the model/training model from the Logistic Regression algorithm have a precision value of 0.79, a recall value of 0.83, an f1 value of 0.83, and a ROC-AUC value of 0.81 after getting five folds of cross-validation.

```
              precision    recall  f1-score   support

           0       0.97      0.79      0.87       967
           1       0.14      0.62      0.23        55

    accuracy                           0.78      1022
   macro avg       0.56      0.70      0.55      1022
weighted avg       0.93      0.78      0.84      1022

ROC AUC score: 0.815

--------------------------------------------------

Cross-validation scores with 5 folds:

ROC AUC: 0.882
precision: 0.79
recall: 0.83
f1: 0.81
```

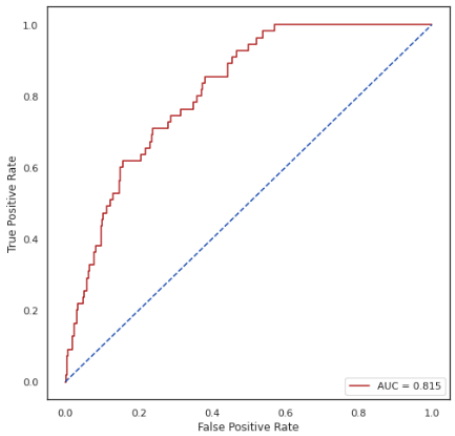

Figure 15. Logistic Regression Result Detail

3.3.2 Random Forest Classifier

The results of making the model/training model from the Random Forest algorithm have a precision value of 0.78, a recall value of 0.9, an f1 value of 0.84, and a ROC-AUC value of 0.921 after getting five folds of cross-validation.

	precision	recall	f1-score	support
0	0.98	0.76	0.86	967
1	0.14	0.71	0.24	55
accuracy			0.76	1022
macro avg	0.56	0.73	0.55	1022
weighted avg	0.93	0.76	0.82	1022

ROC AUC score: 0.829

Cross-validation scores with 5 folds:

ROC AUC: 0.921
precision: 0.78
recall: 0.9
f1: 0.84

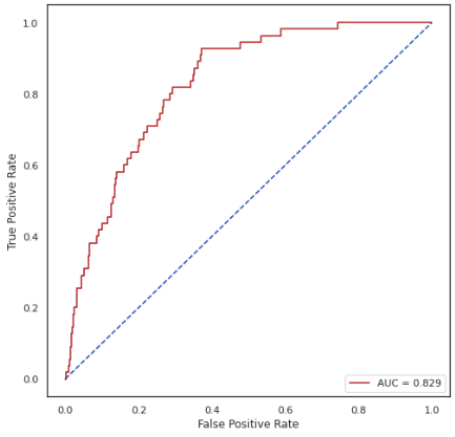

Figure 16. Random Forest Result Detail

3.3.3 Extreme Gradient Boosting

The results of making the model/training model from the XGBoost algorithm have a precision value of 0.9, a recall value of 0.95, an f1 value of 0.92, and a ROC-AUC value of 0.978 after getting five folds of cross-validation.

	precision	recall	f1-score	support
0	0.96	0.89	0.93	967
1	0.18	0.42	0.25	55
accuracy			0.87	1022
macro avg	0.57	0.66	0.59	1022
weighted avg	0.92	0.87	0.89	1022

ROC AUC score: 0.806

Cross-validation scores with 5 folds:

ROC AUC: 0.978
precision: 0.9
recall: 0.95
f1: 0.92

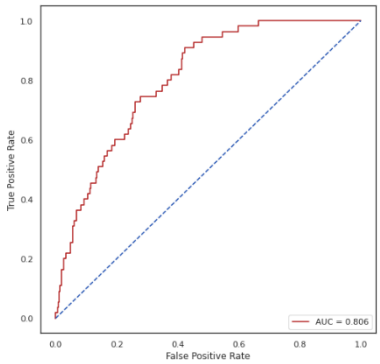

Figure 17. XGBoost Result Detail

4. CONCLUSSION

With the influence of technology in the medical world, stroke can now be predicted using Machine Learning methods based on classification algorithms, so that with predetermined medical parameters, the Machine Learning model that has been created can predict whether a person has a stroke. After making three algorithms that are generally used to carry out the classification process, the results are as follows, Logistic Regression with a precision value of 0.79, a recall value of 0.83, an f1 value of 0.83, and a ROC-AUC value of 0.81 after receiving five folds of cross-validation; Random Forest with a precision value of 0.78, a recall value of 0.9, an f1 value of 0.84 and a ROC-AUC value of 0.921 after getting five folds of cross-validation; and XGBoost with a precision value of 0.79, a recall value of 0.83, an f1 value of 0.83 and a ROC-AUC value of 0.81 after getting five folds of cross-validation. It can be concluded that the best performing algorithm in this case study is XGBoost. This algorithm is suitable for use with pipelines to make real-time predictions.

BIBLIOGRAPHY

[1] C. M. Stinear, C. E. Lang, S. Zeiler, and W. D. Byblow, “Advances and challenges in stroke rehabilitation,” *The Lancet Neurology*, vol. 19, no. 4, pp. 348–360, Apr. 2020, doi: 10.1016/S1474-4422(19)30415-6.

[2] World Health Organization, “The Top 10 Causes of Death,” Dec. 09, 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed Jun. 04, 2022).

[3] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Mar. 2018, pp. 1275–1278. doi: 10.1109/ICECA.2018.8474922.

[4] P. Chantamit-o-pas and M. Goyal, "Long Short-Term Memory Recurrent Neural Network for Stroke Prediction," 2018, pp. 312–323. doi: 10.1007/978-3-319-96136-1_25.

[5] Y. Gu, H. Shen, G. Bai, T. Wang, H. Tong, and Y. Hu, "Incentivizing Multimedia Data Acquisition for Machine Learning System," 2018, pp. 142–158. doi: 10.1007/978-3-030-05057-3_11.

[6] K. Li, "High speed linear array CCD image data acquisition system based on machine learning," 2020. doi: 10.4108/eai.27-8-2020.2294702.

[7] Fedesoriano, "Stroke Prediction Dataset," *Kaggle*, Jun. 2021. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed Jun. 04, 2022).

[8] I. F. Ilyas and X. Chu, "Machine learning and probabilistic data cleaning," in *Data Cleaning*, Association for Computing Machinery, 2019. doi: 10.1145/3310205.3310213.

[9] I. F. Ilyas and T. Rekatsinas, "Machine Learning and Data Cleaning: Which Serves the Other?," *Journal of Data and Information Quality*, Mar. 2022, doi: 10.1145/3506712.

[10] J. Tummers, C. Catal, H. Tobi, B. Tekinerdogan, and G. Leusink, "Coronaviruses and people with intellectual disability: an exploratory data analysis," *Journal of Intellectual Disability Research*, vol. 64, no. 7, pp. 475–481, Jul. 2020, doi: 10.1111/jir.12730.

[11] A. Shabbir, M. Shabbir, A. R. Javed, M. Rizwan, C. Iwendi, and C. Chakraborty, "Exploratory data analysis, classification, comparative analysis, case severity detection, and internet of things in COVID-19 telemonitoring for smart hospitals," *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–28, Feb. 2022, doi: 10.1080/0952813X.2021.1960634.

[12] D. Eilertz, M. Mitterer, and J. M. Buescher, "automRm: An R Package for Fully Automatic LC-QQQ-MS Data Preprocessing Powered by Machine Learning," *Analytical Chemistry*, vol. 94, no. 16, pp. 6163–6171, Apr. 2022, doi: 10.1021/acs.analchem.1c05224.

[13] V. Gómez-Escalonilla, P. Martínez-Santos, and M. Martín-Loeches, "Preprocessing approaches in machine-learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali," *Hydrology and Earth System Sciences*, vol. 26, no. 2, pp. 221–243, Jan. 2022, doi: 10.5194/hess-26-221-2022.

[14] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[15] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[16] C. M. Yeşilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," *Chaos, Solitons & Fractals*, vol. 140, p. 110210, Nov. 2020, doi: 10.1016/j.chaos.2020.110210.

[17] S. H. Shah, Y. Angel, R. Houborg, S. Ali, and M. F. McCabe, "A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat," *Remote Sensing*, vol. 11, no. 8, p. 920, Apr. 2019, doi: 10.3390/rs11080920.

[18] Y. Jia *et al.*, "GNSS-R Soil Moisture Retrieval Based on a XGboost Machine Learning Aided Method: Performance and Validation," *Remote Sensing*, vol. 11, no. 14, p. 1655, Jul. 2019, doi: 10.3390/rs11141655.

[19] I. L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," in *2019 Wireless Days (WD)*, Apr. 2019, pp. 1–6. doi: 10.1109/WD.2019.8734193.