Early Detection of Cardiovascular Disease Using Logistic Regression

Mr. Abhay Baliyan

***Abstract*—Cardiovascular diseases (CVDs) are a major cause of death worldwide, which makes early diagnosis and risk stratification critically important. With the increasing availability of structured clinical data, machine learning can be used to support clinicians in identifying high-risk patients at an early stage. This work presents a predictive model based on Logistic Regression for detecting the presence of heart disease using standard medical attributes such as age, resting blood pressure, cholesterol, chest pain type, maximum heart rate, and other diagnostic indicators. The dataset, obtained from Kaggle, contains 303 patient records with 14 well-defined clinical features and required minimal preprocessing due to its clean structure. Exploratory data analysis was used to understand feature distributions and correlations before splitting the data into training and testing sets using a 60:40 ratio. Logistic Regression was chosen for its interpretability, computational efficiency, and widespread acceptance in clinical decision support systems. The trained model achieved an accuracy of 85.12%, precision of 84.32%, recall of 89.23%, and an F1-score of 86.71%. These results indicate strong predictive performance, especially in recall, which is crucial for minimizing missed diagnoses in medical applications. Overall, the study highlights how interpretable machine learning models can be applied to early cardiovascular risk assessment and suggests that such models can be integrated into clinical workflows to enhance decision-making and improve patient outcomes.**

Keywords— Cardiovascular Disease, Healthcare Analytics, Logistic Regression, Machine Learning, Predictive Modeling, Heart Disease Classification, Clinical Decision Support, Early Diagnosis.

I. INTRODUCTION

Cardiovascular diseases (CVDs) are responsible for millions of deaths every year and remain one of the most serious public health challenges globally [1], [2], [3]. According to global health statistics, heart disease accounts for a significant proportion of premature deaths, hospitalizations, and long-term disability [2], [4], [5]. Because of this high burden, early risk detection and timely intervention are critical to improving survival rates and reducing healthcare costs.

Traditionally, the diagnosis and risk assessment of heart disease depend on clinical expertise, electrocardiograms, laboratory test results, and imaging reports [6], [7], [8]. While physicians are trained to interpret these indicators, manual assessment can be subjective, time-consuming, and prone to variability across practitioners. With the growth of electronic health records and structured datasets, there is a clear opportunity to support medical decision-making through data-driven methods.

Machine learning (ML) provides a set of tools that can uncover patterns in historical patient data and generate predictive models [9], [10], [11]. These models can be used as decision support systems to help clinicians identify patients at higher risk of disease, prioritize further testing, and optimize resource allocation [12], [13]. This study focuses on using a Logistic Regression model to predict heart disease from standard clinical features. The objective is to investigate whether a simple and interpretable model can deliver strong predictive performance while remaining easy to explain and trust in a healthcare environment.

II. LITERATURE REVIEW

Over the last decade, there has been considerable interest in applying machine learning to medical diagnosis, especially for cardiovascular conditions [9], [11], [14]. A range of algorithms has been explored, including classical statistical models and more complex non-linear approaches.

Several studies have used Decision Trees and Random Forests to classify heart disease based on demographic and clinical attributes [15], [16]. These methods are capable of modeling non-linear relationships and can provide some degree of interpretability through feature importance scores [16], [17]. Support Vector Machines (SVMs) have also been popular due to their strong performance on high-dimensional datasets and their ability to handle complex decision boundaries [18], [19].

Neural Networks and deep learning models have been employed as well, especially when larger datasets or imaging data are involved [11], [20]. These models can capture intricate patterns but are often criticized for being “black boxes” because their internal decision processes are not transparent [21], [22]. In critical domains like healthcare, lack of interpretability can be a barrier to clinical adoption, as doctors and hospital

administrators may be reluctant to rely on systems they cannot easily understand or justify [12], [21], [22].

Logistic Regression, on the other hand, has a long history in medical statistics [23], [24]. It is widely used for binary classification problems such as disease vs. no disease and is preferred in many clinical studies due to its simplicity, interpretability, and direct probabilistic output [22], [24]. For heart disease prediction in particular, earlier work such as that based on the UCI Heart Disease dataset has demonstrated that Logistic Regression can achieve competitive performance compared to more complex models, especially when the dataset is structured and well-curated [25], [26].

Building on this body of work, the present study focuses on a well-known heart disease dataset and uses Logistic Regression to create a baseline yet reliable model. The emphasis is on balancing predictive performance with interpretability so that the resulting model can be realistically considered for decision support in clinical settings.

III. METHODOLOGY

A. Overview

The overall approach followed in this study consists of several stages: loading and inspecting the dataset, performing basic data preprocessing, conducting exploratory data analysis (EDA), selecting a suitable model, training and testing the model, and finally evaluating performance using multiple metrics. The workflow is designed to follow best practices in machine learning while keeping the process transparent and reproducible.

B. Dataset Description

The dataset used in this project was obtained from Kaggle and is widely recognized as a benchmark heart disease dataset [27], [28]. It contains 303 records, each representing a single patient, and 14 attributes including demographic and clinical variables [25], [27]. The target variable, usually referred to as target, indicates whether the patient has heart disease (1) or not (0).

Table I lists the medical attributes used in the study along with their dataset notations. These include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and a categorical variable related to thalassemia.

Table I

Description of Medical Attributes Used in the Heart Disease Dataset

SR. NO.	Attribute	Denotation
1	Age	age
2	Sex	sex
3	chest pain type (4 values)	Cp
4	resting blood pressure	trestbps
5	serum cholestoral (in mg/dl)	chol
6	fasting blood sugar (> 120 mg/dl)	Fbs
7	resting electrocardiographic results (values 0,1,2)	restecg
8	maximum heart rate achieved	thalach
9	exercise induced angina	exang
10	oldpeak (ST depression induced by exercise relative to rest)	oldpeak
11	the slope of the peak exercise ST segment	slope
12	number of major vessels (0-3) colored by fluoroscopy	ca
13	thal: 0 = normal; 1 = fixed defect; 2 = reversable defect	thal
14	target (1 if person has disease, else 0)	target

These features are commonly used by clinicians when diagnosing cardiovascular conditions, which makes the dataset well-suited for building a predictive model that aligns with real-world medical practice [6], [25].

C. Data Preprocessing

One of the advantages of this dataset is that it is already relatively clean. A check for missing values showed no null entries across the attributes. Basic descriptive statistics were computed to understand the central tendency and spread of each feature. No extreme anomalies or impossible values were detected.

Since all features were numerical or encoded as numeric categories, there was no need for additional categorical encoding [29], [30]. For this baseline model, advanced preprocessing steps such as normalization or standardization were not strictly required to make Logistic Regression converge, but they can be explored in future work for potential performance gains.

D. Exploratory Data Analysis

Exploratory data analysis was performed to better understand the relationships between features and the target variable. The first step was to visualize the distribution of disease vs. no disease in the dataset. **Fig. 1** illustrates the class distribution, showing how many patients are labeled as having heart disease compared to those who do not. This helps in assessing whether the dataset is balanced or skewed toward a particular class.

Scatter plots and count plots were used to explore how variables such as age, cholesterol level, and chest pain type relate to the presence of heart disease. For example, a scatter plot of age vs. cholesterol, colored by disease status, provides insight into whether higher cholesterol is more prevalent in patients with heart disease.

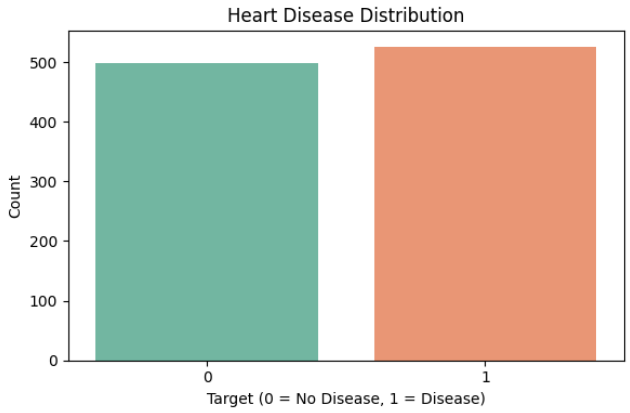

Fig. 1. Distribution of heart disease cases in the dataset

A correlation matrix was also computed to analyze how each feature correlates with the others and with the target variable. The correlation matrix, visualized as a heatmap in **Fig. 2**, highlights which attributes are strongly related and can guide future feature selection or dimensionality reduction efforts.

Fig. 2. Correlation matrix of features used in the model

E. Model Selection

Given the goal of creating a clinically useful and interpretable model, Logistic Regression was selected. Logistic Regression is a linear model that estimates the probability of a binary outcome given a set of input features [23], [24]. The output is a probability between 0 and 1, which can be interpreted as the estimated likelihood that a patient has heart disease.

Another advantage is that Logistic Regression provides coefficients for each feature, which can be analyzed to understand whether a particular feature increases or decreases the probability of disease [22], [24]. This is particularly important in the medical domain where interpretability and explanation are essential [12], [21], [22].

F. Model Training

The dataset was split into training and testing sets using a 60:40 ratio via the train_test_split function[29], [31]. This means that 60% of the data was used to fit the model, and the remaining 40% was held out for evaluating performance on previously unseen examples.

The Logistic Regression model was created with a maximum iteration value (max_iter) of 1500 to ensure convergence [24], [29]. The model was trained using the training subset, where it learned coefficients corresponding to each clinical feature.

G. Evaluation Metrics and Protocol

To evaluate performance, several metrics were computed on the test set: accuracy, precision, recall, and F1-score[32], [33].

- **Accuracy** measures the overall proportion of correct predictions [32], [33].
- **Precision** indicates what fraction of patients predicted as having heart disease actually do have it [32], [33].
- **Recall** (sensitivity) measures how many of the actual disease cases were successfully identified [32], [33].
- **F1-score** is the harmonic mean of precision and recall, providing a single measure that balances both [32], [33].

A confusion matrix was also generated to visualize true positives, true negatives, false positives, and false negatives. This is presented in **Fig. 3**. In medical contexts, recall is especially important because failing to identify a patient who truly has heart disease (a false negative) can have serious consequences[22], [32], [33].

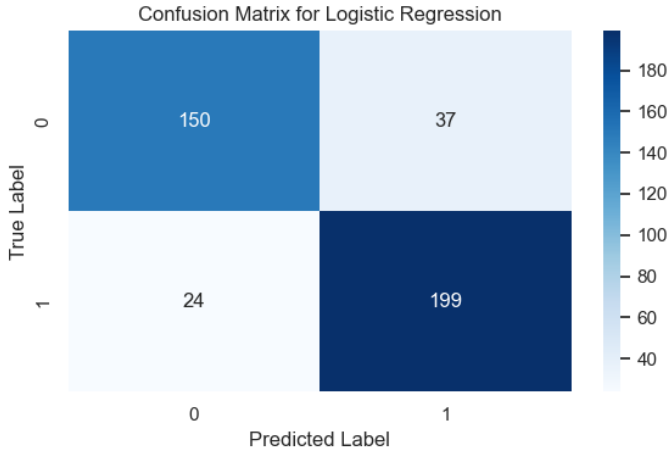

Fig. 3. Confusion matrix for the logistic regression model

Optionally, a Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) can also be computed to further analyze the trade-off between true positive rate and false positive rate [34], [35].

IV. RESULTS

A. Quantitative Performance

After training the model, predictions were made on the test set and the evaluation metrics were calculated. The model achieved an accuracy of 85.12%, which indicates that roughly 85 out of

100 patients were correctly classified as either having or not having heart disease.

The precision score was 84.32%, meaning that most of the patients predicted as positive (with heart disease) were truly positive [32]. The recall score reached 89.23%, which is particularly encouraging because it shows that a large portion of actual heart disease cases were correctly identified by the model [32], [33]. The resulting F1-score of 86.71% confirms a strong balance between precision and recall [32], [33].

These values are summarized in **Table II**.

Table II

Performance Metrics of the Logistic Regression Model

Metric	Value (%)
Accuracy	85.12
Precision	84.32
Recall	89.23
F1 Score	86.71

B. Confusion Matrix Analysis

The confusion matrix in **Figure 3** provides more detailed insight into how the model performs across the positive and negative classes. A high number of true positives and true negatives indicates that the model generally assigns correct labels. The relatively low count of false negatives is particularly important for a diagnostic tool, as missed disease cases can lead to delayed treatment.

C. Feature Influence

While this study did not focus on a full statistical analysis of model coefficients, Logistic Regression inherently provides information about how each feature impacts the prediction [22], [24]. Positive coefficients indicate that an increase in that feature value is associated with a higher likelihood of heart disease, and negative coefficients indicate the opposite [23], [24]. For example, higher values of chest pain type or ST depression (oldpeak) may be associated with increased risk, while certain patterns in heart rate may be protective or neutral. This capacity to interpret feature influence is a key reason for choosing Logistic Regression in this context [12], [22].

V. DISCUSSION

The results demonstrate that even a relatively simple model like Logistic Regression can achieve strong performance in heart disease prediction when applied to a well-structured dataset [25], [26]. Without heavy preprocessing, feature engineering, or model tuning, the model still reached an accuracy above 85% and a high recall value.

From a clinical perspective, such a model could be used as a preliminary screening tool [12], [13]. For example, it could run in the background of an electronic health record system and flag patients whose clinical measurements suggest an elevated risk of heart disease [12], [13], [36]. Clinicians could then review these recommendations and decide whether additional testing or specialist referral is warranted.

Another important aspect is interpretability. Unlike complex black-box models, Logistic Regression makes it easier to understand why a particular prediction was made [21], [22]. This aligns well with the expectations of healthcare professionals who need to justify decisions to patients, peers, and regulatory bodies [12], [22].

However, it is also important to recognize that this study is based on a single dataset with a relatively small sample size (303 patients). Real-world clinical data may be noisier, more diverse, and may contain missing or inconsistent entries [10], [37]. Therefore, while the model shows promise, it should not be considered a fully validated diagnostic tool without further testing and external validation [11], [37].

VI. LIMITATIONS

Several limitations should be noted:

A. **Dataset Size and Diversity:** The dataset is relatively small and may not fully capture the variability in real-world patient populations across different regions, ethnicities, and healthcare systems [10], [37].

B. Single Dataset: The study relies on a single public dataset. Cross-dataset validation or external validation with hospital data would be necessary before clinical deployment [11], [37].

C. Limited Feature Set: Only the attributes provided in the dataset were used. Important risk factors such as family history, smoking status, physical activity, diet, and medication use were not included [6], [7].

D. Basic Model Configuration: The Logistic Regression model used in this study employs default hyperparameters except for the maximum number of iterations. More extensive hyperparameter tuning might yield slightly better results.

E. No Comparison with Other Models in This Implementation: Although literature suggests that other models like Random Forest, SVM, and XGBoost can perform competitively or better [16], [26], [38], a direct head-to-head comparison was not performed here and remains open for future work.

VII. FUTURE WORK

There are several directions in which this work can be extended:

A. Model Comparison: Future studies can compare Logistic Regression with other algorithms such as Random Forest, Gradient Boosting, Support Vector Machines, and Neural Networks to determine performance trade-offs between accuracy and interpretability [16], [19], [20], [26], [38], [39].

B. Hyperparameter Tuning and Cross-Validation: Techniques such as grid search, randomized search, and k-fold

cross-validation can be applied to optimize model parameters and obtain more reliable performance estimates [40], [41].

C. Feature Engineering: Additional derived features, such as risk scores or interaction terms between variables, may further improve predictive power [17], [31].

D. Handling Larger and Real-World Datasets: Applying the model to larger and more diverse clinical datasets from hospitals or health systems would help validate its robustness in practical scenarios [10], [11], [37].

E. Deployment as a Clinical Tool: A logical next step would be to integrate the model into a simple web or desktop application for clinicians, or to embed it in an electronic health record system as a background risk assessment tool [12], [13], [42].

VIII. CONCLUSION

This study developed and evaluated a Logistic Regression–based model for early detection of heart disease using a structured clinical dataset. With minimal preprocessing and a straightforward modeling approach, the model achieved strong results, including an accuracy of 85.12% and an F1-score of 86.71% [32]. These metrics, combined with high recall, indicate that the model can effectively identify patients at risk of cardiovascular disease.

The key advantage of the approach lies not only in its performance but also in its interpretability [21], [22]. Logistic Regression provides clear insight into how each clinical attribute contributes to the prediction, which is an essential requirement in healthcare applications [12], [22], [24]. While there are limitations related to dataset size and the absence of external validation, the results suggest that interpretable machine learning models can play a meaningful role in supporting clinicians and improving early diagnosis.

Overall, this work demonstrates that even relatively simple models, when applied thoughtfully, can offer valuable support for cardiovascular risk assessment and can serve as a practical baseline for more advanced research and development [11], [22], [42].

ACKNOWLEDGMENT

I would like to thank the open-source community and Kaggle for providing accessible datasets that make research and experimentation possible. Appreciation is also extended to educators, researchers, and practitioners in the fields of data science and healthcare whose contributions inspired this work.

REFERENCES

[1] World Health Organization, “Cardiovascular diseases (CVDs),” *WHO Fact Sheet*, 2023, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, and others, “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” *Circulation*, vol. 147, no. 8, pp. e93–e621, 2023, doi: 10.1161/CIR.0000000000001123.

[3] World Heart Federation, “World Heart Report 2023: Confronting the World’s Number One Killer,” Geneva, Switzerland, 2023. [Online]. Available: https://world-heart-federation.org

[4] Centers for Disease Control and Prevention, “Heart Disease Facts,” Atlanta, GA, 2023. [Online]. Available: https://www.cdc.gov/heart-disease/data-research/facts-stats/

[5] A. Timmis, P. Vardas, N. Townsend, and others, “European Society of Cardiology: The 2023 Atlas of Cardiovascular Disease Statistics,” *Eur Heart J*, vol. 45, no. 34, pp. 3085–3164, 2024, doi: 10.1093/eurheartj/ehae466.

[6] R. B. D’Agostino, R. S. Vasan, M. J. Pencina, and others, “General cardiovascular risk profile for use in primary care: the Framingham Heart Study,” *Circulation*, vol. 117, no. 6, pp. 743–753, 2008, doi: 10.1161/CIRCULATIONAHA.107.699579.

[7] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, and others, “2013 ACC/AHA guideline on the assessment of cardiovascular risk,” *Circulation*, vol. 129, no. 25 Suppl 2, pp. S49–S73, 2014, doi: 10.1161/01.cir.0000437741.48606.98.

[8] P. W. F. Wilson, R. B. D’Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, “Prediction of coronary heart disease using risk factor categories,” *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998, doi: 10.1161/01.CIR.97.18.1837.

[9] F. Jiang, Y. Jiang, H. Zhi, and others, “Artificial intelligence in healthcare: past, present and future,” *Stroke Vasc Neurol*, vol. 2, no. 4, pp. 230–243, 2017, doi: 10.1136/svn-2017-000101.

[10] A. L. Beam and I. S. Kohane, “Big Data and Machine Learning in Health Care,” *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018, doi: 10.1001/jama.2017.18391.

[11] A. Rajkomar, J. Dean, and I. Kohane, “Machine Learning in Medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/NEJMra1814259.

[12] E. H. Shortliffe and M. J. Sepúlveda, “Clinical Decision Support in the Era of Artificial Intelligence,” *JAMA*, vol. 320, no. 21, pp. 2199–2200, 2018, doi: 10.1001/jama.2018.17163.

[13] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An

overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digit Med*, vol. 3, no. 1, pp. 1–10, 2020, doi: 10.1038/s41746-020-0221-y.

[14] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci Rep*, vol. 7, no. 1, p. 12967, 2017, doi: 10.1038/s41598-017-13314-1.

[15] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of Heart Disease Using Random Forest and Feature Subset Selection," in *Innovations in Bio-Inspired Computing and Applications*, vol. 424, Springer, 2016, pp. 187–196. doi: 10.1007/978-3-319-28031-8_16.

[16] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[18] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network," in *International Conference on Computer and Communication Technology (ICCCT)*, 2010, pp. 741–745. doi: 10.1109/ICCCT.2010.5640377.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[22] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM, 2015, pp. 1721–1730. doi: 10.1145/2783258.2788613.

[23] D. G. Kleinbaum and M. Klein, *Logistic Regression: A Self-Learning Text*, 3rd ed. New York: Springer, 2010.

[24] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. in Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, 2013. doi: 10.1002/9781118548387.

[25] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989, doi: 10.1016/0002-9149(89)90524-9.

[26] T. Choudhury, V. Kumar, R. Gupta, and N. Pradhan, "A Comparative Analysis of Machine Learning Techniques for Heart Disease Prediction," in *2021 IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 167–172. doi: 10.1109/ICCCIS51004.2021.9397207.

[27] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease Dataset," 1989, *University of California, Irvine*. doi: 10.24432/C52P4X.

[28] Kaggle Community, "UCI Heart Disease Data," 2020, *Kaggle Inc.* [Online]. Available: https://www.kaggle.com/datasets/ronitf/heart-disease-uci

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[31] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2013.

[32] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[33] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[34] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.

[35] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[36] Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *New England Journal of Medicine*, vol.

375, no. 13, pp. 1216–1219, 2016, doi: 10.1056/NEJMp1606181.

[37] L. Wynants, B. Van Calster, G. S. Collins, and others, "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal," *BMJ*, vol. 369, p. m1328, 2020, doi: 10.1136/bmj.m1328.

[38] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[39] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann Stat*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[40] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[41] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1995, pp. 1137–1143.

[42] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat Med*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.