Comparative Study of Decision Tree, AdaBoost, Random Forest, Naïve Bayes, KNN, and Perceptron for Heart Disease Prediction

Mohammad Maydanchi[1*], Armin Ziaei[2], Mina Basiri[3], Alireza Norouzi Azad[4], Shaheen Pouya[1], Mehrbod Ziaei[5], Fatemeh Haji[6], Saman Sargolzaei[7]

[1]*Department of Industrial and Systems Engineering, Auburn University,* AL, USA
[2]*Department of Engineering and Computer Science, The University of Texas at Dallas,* TX, USA
[3]*Department of Industrial and Systems Engineering, Tarbiat Modares University,* Tehran, Iran
[4]*Department of Engineering Science, University of Tehran,* Tehran, Iran
[5]*Department of Electrical Engineering, University of Science and Culture,* Tehran, Iran
[6]*Department of Computer Engineering, Ferdowsi University of Mashhad,* Mashhad, Iran
[7]*Department of Engineering, The University of Tennessee at Martin,* TN, USA

***Abstract*— Globally, cardiovascular diseases (CVD) are estimated to account for more than 32% of all deaths. Consequently, CVD has become a global health problem, and timely diagnosis is essential (WHO, 2021). Screening for risk factors accelerates the diagnosis and management of CVD, resulting in a more effective and rapid response, reducing the risk of death. This article compares six classification models, AdaBoost, Random Forest, Decision Tree, KNN, Naive Bayes, and Perceptron, to predict CVD symptoms. Based on CDC data collected from Kaggle, classification models were compared with the approach of examining effective factors to predict heart disease. Since the data set was imbalanced, the study performance was measured by AUC and F_1-score in Class 1, which is the critical class in this dataset. AdaBoost is found to have the highest AUC and F_1-score, respectively, of 0.828 and 0.37, while Decision Tree has the lowest AUC of 0.595 and F_1-score of 0.25.**

Keywords—Cardiovascular diseases, Classification models, AdaBoost, Random Forest, Decision Tree, KNN, Naïve Bayes, Perceptron.

I. INTRODUCTION AND LITERATURE REVIEW

It is consistently reported that cardiovascular disease (CVD), also known as heart disease, is one of the greatest threats to human lives [1] - [3]. The World Health Organization (WHO) reports that around 18 million people worldwide suffer from heart failure yearly, accounting for 32% of all casualties. For the foreseeable future, this rate does not appear to be decreasing. Patients with small hearts should be seen regularly by their specialists. According to Abdoli et al., choosing an effective scheduling policy is critical in improving clinic performance and patient satisfaction [4]. Aside from this, there are numerous conditions in which patients are unaware they are suffering from heart failure (silent heart attack). In addition to the people mentioned above, approximately 20% of silent heart attack cases have been reported [3], [5], [6].

WHO reports indicate that road injuries are the only cause of fatalities unrelated to health conditions or diseases. The number of deaths caused by traffic violations yearly is not even comparable to those caused by heart disease. Fatalities from ischemic heart disease are more than seven times greater than those from road accidents [7, 8]. The application of lean principles improves the performance of emergency departments, which play a significant role in the survival of heart patients [9]. Cardiovascular diseases are associated with numerous difficulties. Several difficulties have been identified, including substantial morbidity, financial concerns, and general population policies. It is estimated that approximately 6.5 million Americans are affected by heart disease (together with 5.7 million Alzheimer's patients who also suffer from heart conditions) [10, 11]. Islamipour and Nobari have developed a multi-objective model to create a sustainable blood supply chain, which includes multiple donors, collection centers, distribution centers, and hospitals at various levels. [12]. In addition, Vincent et al. [13] report that by 2050, the elderly population in the United States will number over 88 million. Given the high risk of heart disease for this group and the financial consequences caused by their medical needs, extra consideration should be given to heart diseases [14], [15].

Currently, there are several types of heart disease, some of which are more common and some of which might cause more severe damage than others. Some known types are coronary artery disease, heart arrhythmias, heart failure, heart valve disease, pericardial disease, cardiomyopathy, and congenital heart disease [16]. Although the highest correlating factor to the heart failure rate would be age, there are also some significant risks causing cardiovascular disease or enhancing the heart failure rate. These factors include age, being inactive, diabetes, genetics, family history, high low-density lipoprotein or LDL, known as bad cholesterol, "blood pressure, smoking, and high level of stress [17]. Even though numerous methods or campaigns exist to reduce CVD, control its causality, and model

*Corresponding author, mzm0181@auburn.edu.

978-1-6654-7611-9/23/$31.00 ©2023 IEEE

its effects and causes, this phenomenon will continue to pose a high risk to future generations [18], [19]. On the other hand, due to the significant changes in everyday life due to the COVID-19 pandemic, remote working trends, particularly in cities, and global health programs to prevent CVD, there are some positive signs for the future [20]– [22]. In addition, many countries are running cost-related programs to reduce 30% of CVD mortality [17].

The field of artificial intelligence, which may be loosely described as the ability of a computer to reproduce intelligent human behavior, is the overarching field within which machine learning falls. Artificial intelligence systems are used to complete complex tasks that can be compared to the way people solve problems. In general, it refers to machines capable of recognizing visual scenes, comprehending natural language texts, or executing physical actions. A machine learning algorithm begins with data. Data can take the form of numbers, time-series data, images, or text, as well as financial transactions, videos of people, musical instruments, or phone records. "Heart disease" refers to various illnesses affecting the heart. Unfortunately, we often lack sufficient information in many cases. Sepehriyar et al. Proposed a method to deal with imperfect or incomplete data [23].

Machine learning is used to predict diseases based on the symptoms that patients or users provide. As an input, the user enters the symptoms they are experiencing, and the system generates an output representing the likelihood of a particular disease occurring. Effective analysis of medical data is essential for early diagnosis and treatment of illness, and this has become more critical with the growth of biological and healthcare data [24]. Heart disease can be detected earlier through enhanced diagnostics and high-risk patients utilizing a prediction model, reducing fatality rates and improving the decision-making process for future treatment and prevention [25]. Various ailments may be diagnosed, detected, or predicted using machine learning in medicine. This research aims to provide doctors with a method for identifying cardiac issues as fast as possible. This way, patients can receive appropriate therapy while minimizing adverse effects [26]. There are many publicly available sources for accessing patients' medical records, and studies can be carried out so that a variety of computer technologies can be utilized for making an accurate diagnosis of patients and detecting this illness to stop it from progressing to a point where it can be fatal. As is well known, machine learning and artificial intelligence are playing a significant role in the field of medicine in the present day. Different machine learning and deep learning models may be used to diagnose illnesses, categorize outcomes, or predict their outcomes. The use of machine learning models makes it possible to analyze genetic data in a comprehensive manner. It is possible to train models to make reliable pandemic forecasts, and medical records can also be altered and studied more thoroughly to make more accurate predictions. At work, the schedule must also be observed. Mohammedadian et al. propose a new model to optimize nurses' scheduling in emergency rooms based on their priorities. The efficiency of workload and server scheduling significantly impact the quality of care [27].

The classification and prediction of heart disease have been the subject of extensive research, and many machine-learning models have been developed. Using a machine learning algorithm called CART, which stands for Classification and Regression, Melillo et al. [28] created an automatic classifier that can distinguish between patients at high risk of congestive heart failure and those at low risk of the condition. The sensitivity of the classifier was found to be 93.3%, while the specificity was found to be 63.5%. To enhance the method's performance, Liao et al. [29] propose utilizing the electrocardiogram (ECG) method. Deep neural networks are used in this method to select the most critical information for employees, and then these features are incorporated into the system.

In addition, Guidi et al. [30] have developed a clinical decision support system to detect cardiac failure early to prevent it. They attempted to evaluate and contrast a variety of machine learning models and deep learning models, particularly neural networks, using techniques such as CART [31], support vector machines, and random forests. According to [32], a deep convolutional neuro-fuzzy network was employed to identify acute lymphoblastic leukemia (ALL) from microscopic cell images. On average, this model was able to detect acute lymphoblastic leukemia with an accuracy of 97.31%. Based on unstructured clinical notes, Zhang et al. [33] found that the rule-based method combined with natural language processing led to an accuracy rate of 93.37 % when determining the NYHA HF (New York Heart Association Functional Classification) class. Parthiban and Srivatsa [33, 34] used support vector machines to identify patients who already have diabetes and to predict heart disease. These techniques achieved an accuracy rate of 94.60 %, and the features used were those most commonly associated with diabetes, such as blood sugar level, age, and blood pressure. In this paper, we aim to compare the performance of AdaBoost, Random Forest, Decision Tree, KNN, Naive Bayes, and Perceptron using data from the Centers for Disease Control and Prevention (CDC). Firstly, we describe the models, the dataset, and its characteristics in the materials and methods section. In the next step, we compare the models, and in the final stage, we draw conclusions about the work and discuss possible directions for future research.

II. MATERIALS AND METHODS

Kaggle originally obtained the dataset from CDC, which was also provided by Kaggle (www.kaggle.com, 2020). In collaboration with all U.S. states, participating U.S. territories, and the Centers for Disease Control and Prevention (CDC), the Behavioral Risk Factor Surveillance System (BRFSS) was developed [35]. Based on the CDC's 2020 annual health survey of 400,000 adults, this is information about their health status. To better predict heart disease, machine learning provides a better understanding of the factors that influence it. Initially, the data were preprocessed to extract attributes that effectively predict the occurrence of heart disease. Then, the model utilizes these attributes to estimate whether a person will have heart disease. In this case, the target variable is a binary one. In light of the importance of identifying people with heart disease, data related to patients are divided into two classes. These classes are class 1, patients with heart disease, and class 0, patients without heart disease. Class 1 and class 0 have 27373 and 292422 samples, respectively, indicating that the dataset is imbalanced. Even though class 1 samples comprise only 8.5% of the dataset, it is evident that the classification of samples belonging to class 1 is more critical than that of class 0. There are two types of attributes, categorical and continuous, which are listed in Table I.

TABLE I. FEATURES OF THE DATASET

Feature	Type	Range of Value
BMI	Continues	Float [1-9999]
Physical Health	Continues	Number of days [1-30]
Mental Health	Continues	Number of days [1-30]
Sleep Time	Continues	Number of hours [1-24]
Heart Disease	Categorical	Yes/ No
Smoking	Categorical	Yes/ No
Alcohol Drinking	Categorical	Yes/ No
Stroke	Categorical	Yes/ No
Difficulty in walking	Categorical	Yes/ No
Sex	Categorical	Male/ Female
Race	Categorical	White/ Black/ Asian/ American Indian or Alaskan Native/ Hispanic/ Other
Diabetic	Categorical	Yes/ No/ No, borderline diabetes/ Yes (during pregnancy)
Physical Activity	Categorical	Yes/ No
General Health	Categorical	Excellent/ Very good/ Good/ Fair/ Poor
Asthma	Categorical	Yes/ No
Kidney Disease	Categorical	Yes/ No
Skin Cancer	Categorical	Yes/ No
Age Category	Categorical	[18-79]/ [80 or older]

As seen, 14 of the 18 attributes are categorical, and only four are continuous. Therefore, we divided the data into a training dataset (80%) and a test dataset (20%) to extract and use patterns in the data and evaluate the models. Since the data sets in this article are imbalanced, the standardized method is employed for scaling. To handle the imbalance problem, we used SMOTE method[42]. In addition, categorical features were encoded to run the model. However, when the number of parameters is large, the model becomes more complex, which may decrease its efficiency. Thus, K-fold cross-validation was used to avoid overfitting. In this paper, the performance of six classification models in our dataset was compared with each other, which are introduced below.

Perceptron is a supervised learning algorithm for binary classifications that enables neurons to simultaneously learn and process elements of the training set [36]. As a simple supervised machine learning algorithm capable of solving classification and regression problems, the K-nearest neighbor (KNN) algorithm was also employed in this study [37]. As a basis for classification algorithms, Naive Bayes is the next algorithm. The Naive Bayes approach assumes that one feature does not depend on another [38]. The purpose of boosting algorithms is to achieve strong learners by combining several weak models. It is used here as an ensemble method, along with AdaBoost, also known as adaptive boosting [39]. This algorithm is based on the concept of ensemble learning and takes predictions from each tree based on the majority of forecasts and predicts the final result [40]. Random forest and decision tree are also supervised learning models. However, unlike a random forest, a decision tree is a type of supervised machine learning that employs a tree-like model of decisions and their possible consequences, in which data is continuously divided based on a particular parameter [41]. Table II presents the parameters for each model.

TABLE II. PARAMETERS OF THE CLASSIFICATION MODELS USED.

Model	Parameter
AdaBoost	Learning rate = 1
Naïve Bayes	Variance smoothing = 1e-9
Random Forest	Criterion: Entropy
Perceptron	Penalty = None
KNN	k = 2
Decision Tree	Criterion: Entropy

These six traditional supervised machine learning methods are selected for this comparison because of their ability to handle independent features and perform prediction tasks with a large number of parameters. The computations are coded in Python 3.10.8 and can be found at https://github.com/orgs/Machine-Learning-Projects1/repositories. The result of this comparison is presented in the following section.

III. RESULTS AND DISCUSSIONS

Throughout this section, we compare the classification strengths of AdaBoost, Random Forest, Decision Tree, KNN, Naïve Bayes, and Perceptron. Table III presents the calculated performance measures. Our Naïve Bayes model had a maximum accuracy of 89% following training, while Perceptron obtained the lowest accuracy of 69%. According to the previous section, each sample's target variable belongs to either class 0 or class 1. Since the number of samples in class 0 (292422) is more than ten times more numerous than the number of samples in class 1, the dataset is moderately imbalanced; consequently, accuracy cannot be used as an effective factor in analyzing the differentiation between classes.

TABLE III. EVALUATION METRICS FOR CLASS 0 AND CLASS 1.

Model	**Accuracy**	**Class 0 (No)**			**Class 1 (Yes)**		
		Precision	Recall	F_1-score	Precision	Recall	F_1-score
AdaBoost	81%	96%	83%	89%	26%	64%	37%
Naïve Bayes	89%	94%	95%	94%	35%	31%	32%
Random Forest	88%	93%	94%	94%	31%	26%	28%
Perceptron	69%	96%	69%	80%	16%	64%	26%
KNN	86%	93%	91%	92%	22%	30%	26%
Decision Tree	86%	93%	91%	92%	22%	27%	25%

In order to accurately identify patients who, have heart disease, it is crucial to classify samples belonging to class 1 rather than class 0. Thus, we compare the performance of models on class 1 based on the F_1-score and ROC curve rather than accuracy to obtain more reliable results.

The precision for class 1 shows the portion of patients who genuinely belong to class 1, and the model classifies them in class 1. Also, the recall number for class 1 in this dataset means the portion of samples in class 1 which are classified correctly. Table III illustrates that Naive Bayes has the highest precision of 35%, and Perceptron reached the lowest precision of 16%. Moreover, AdaBoost and Perceptron achieved the highest recall of 64% and the lowest recall of 26% obtained by Random Forest. However, considering the combination of these two metrics, we see AdaBoost has the highest F_1-score of 37%, whereas Decision Tree has the lowest F_1-score of 25%.

The Receiver Operating Curve (ROC) is another metric used in this study for comparing models. The true positive and false positive rates for a single classifier are calculated and plotted for a ROC curve. The true positive rate is another term for the recall, while the false positive rate is the fraction of the number of samples wrongly classified as class 1 to the number of all the samples that genuinely do not belong to class 1. The ROC curves of six models are shown in Fig. 1. Each model with the highest area under the curve (AUC) performs best. AdaBoost, with an AUC of 0.828, is the best model, and Decision Tree, with an AUC of 0.595, is the worst model among all six for this metric.

IV. CONCLUSIONS

In this study, we used six classification models to predict heart diseases, including Perceptron, KNN, Naive Bayes, AdaBoost, Random Forest, and Decision Tree. In order to compare machine learning models, we reported AUC and F_1-score instead of accuracy because our data set was imbalanced. Due to the significance of class 1, this comparison is limited to this group of data. The AdaBoost model, with an AUC of 0.821 and an F_1-score of 37%, reaches the best results, while the Decision Tree model, with an AUC of 0.588 and an F_1-score of 25%, reaches the worst results.

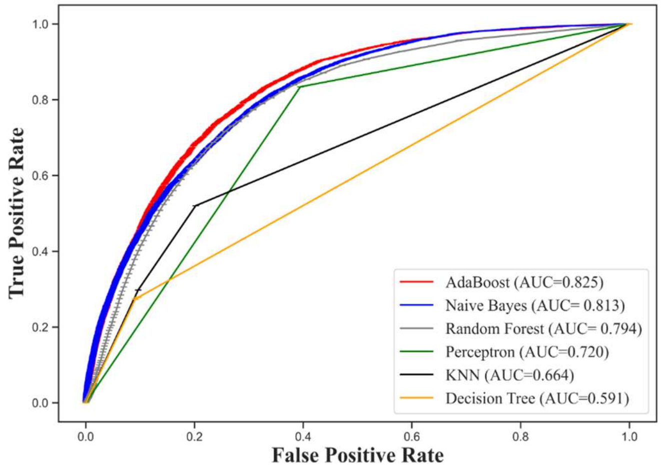

Fig. 1. The ROC curves for the classification.

This research demonstrates that machine learning technology can be utilized as a clinical aid in the diagnosis of cardiovascular illness and can be extremely useful to physicians in situations where a diagnosis is incorrectly made because any error made during a medical diagnosis process could be extremely time-consuming and costly. In addition, more complex models, such as neural networks, may be used to improve future research results.

REFERENCES

[1] WHO, "Cardiovascular diseases (CVDs) (no date) Who.int. Available at:https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Accessed: August 22, 2022)," 2021.

[2] N. S. et al Shah, "Heterogeneous trends in burden of heart disease mortality by subtypes in the United States, 1999-2018: observational analysis of vital statistics Enhanced Reader," doi: 10.1136/bmj.m2688., 2020.

[3] Tsao et al., "Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association," 2022, doi: 10.1161/CIR.0000000000001052.

[4] M. Abdoli, M. Bahadori, R. Ravangard, M. Babaei, and M. Aminjarahi, "Comparing 2 Appointment Scheduling Policies Using Discrete-Event Simulation," Qual. Manag. Health Care, vol. 30, no. 2, Apr. 2021, doi: 10.1097/QMH.0000000000000292.

[5] Cdc, "CDC (2022) Heart disease facts, Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/heartdisease/facts.htm (Accessed: August 22, 2022)," 2022.

[6] S. S. Virani et al., "Heart Disease and Stroke Statistics-2021 Update A Report From the American Heart Association WRITING GROUP MEMBERS On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee," vol. 143, pp. 254–743, 2021, doi: 10.1161/CIR.0000000000000950.

[7] W. H. Organization, "world health organization," Health Topics, 9 December 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Accessed 9 December 2020].

[8] M. Rababah et al., "Data Visualization of Traffic Violations in Maryland, U.S," arXiv [cs. D.B.], Aug. 22, 2022. [Online]. Available: http://arxiv.org/abs/2208.10543

[9] M. Aminjarahi, M. Abdoli, Y. Fadaee, F. Kohan, and S. Shokouhyar, "The Prioritization of Lean Techniques in Emergency Departments Using VIKOR and SAW Approaches," Ethiop. J. Health Sci., vol. 31, no. 2, p. 283, Mar. 2021, doi: 10.4314/ejhs.v31i2.11.

[10] J. T. Vuong et al., "Mortality From Heart Failure and Dementia in the United States: CDC WONDER 1999–2016," J. Card. Fail., vol. 25, no. 2, pp. 125–129, 2019, doi: 10.1016/j.cardfail.2018.11.012.

[11] S. C. Curtin, "Trends in Cancer and Heart Disease Death Rates Among Adults Aged 45-64: United States, 1999-2017," Natl. Vital Stat. Rep., vol. 68, no. 5, pp. 1–9, 2019, [Online]. Available: http://europepmc.org/abstract/MED/32501204

[12] R. Eslamipoor and A. Nobari, "A Reliable and Sustainable Design of Supply Chain in Healthcare Under uncertainty Regarding Environmental Impacts," Journal of Applied Research on Industrial Engineering, vol. 0, Jun. 2022, doi: 10.22105/jarie.2022.335389.1461.

[13] Grayson K VincentVictoria Averil VelkoffU.S. Census Bureau., "The Next Four Decades: The Older Population in the United States : 2010 to 2050. [Washington, D.C.]: [U.S. Dept. of Commerce, Economics and Statistics Administration, U.S. Census Bureau], 2010," 2010.

[14] L. R. Teixeira et al., "The effect of occupational exposure to noise on ischaemic heart disease, stroke and hypertension: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-Related Burden of Disease and Injury," Environ. Int., vol. 154, p. 106387, 2021, doi: 10.1016/j.envint.2021.106387.

[15] S. Barco et al., "Trends in mortality related to pulmonary embolism in the European Region, 2000–15: analysis of vital registration data from the WHO Mortality Database," The Lancet Respiratory Medicine, vol. 8, no. 3, pp. 277–287, 2020, doi: 10.1016/S2213-2600(19)30354-6.

[16] Y. Kokubo and C. Matsumoto, "Hypertension Is a Risk Factor for Several Types of Heart Disease: Review of Prospective Studies," in Hypertension: from basic research to clinical practice, M. S. Islam, Ed. Cham: Springer International Publishing, 2017, pp. 419–426. doi: 10.1007/5584_2016_99.

[17] G. A. Roth et al., "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study," J. Am. Coll. Cardiol., vol. 76, no. 25, pp. 2982–3021, 2020, doi: 10.1016/j.jacc.2020.11.010.

[18] O. Terrada, S. Hamida, B. Cherradi, A. Raihani, and O. Bouattane, "Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease," Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 5, pp. 269–277, 2020.

[19] N. Mohammadi et al., "A randomized trial of an optimism training intervention in patients with heart disease," Gen. Hosp. Psychiatry, vol. 51, pp. 46–53, 2018, doi: 10.1016/j.genhosppsych.2017.12.004.

[20] S. N. M. K. J. Pouya, "Effective factors on virtual communication in projects," The WEI International Academic Conference Proceedings, Antalya, Turkey, Jan. 12, 2014.

[21] S. Pouya, M. Naaranoja, and J. Kantola, "PROJECT STAKEHOLDERS'SUPPOSITION ABOUT PROJECT SUCCESS FACTORS (A COMMUNICATION PERSPECTIVE)," Scientific Committee–Editorial Board, p. 165, 2014.

[22] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, "HDPS: Heart disease prediction system," in 2011 Computing in Cardiology, pp. 557–560, 2011.

[23] A. Sepehriar, R. Eslamipoor, and A. Nobari, "A new mixed fuzzy-LP method for selecting the best supplier using fuzzy group decision making," Neural Comput. Appl., vol. 23, no. 1, pp. 345–352, Aug. 2013, doi: 10.1007/s00521-013-1458-z.

[24] C K Gomathy, Mr. A. Rohith Naidu, "THE PREDICTION OF DISEASE USING MACHINE LEARNING," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 05, no. 10, 2021.

[25] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Comput. Intell. Neurosci., vol. 2021, p. 8387680, Jul. 2021, doi: 10.1155/2021/8387680.

[26] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," J. Healthc. Eng., vol. 2022, 2022, doi: 10.1155/2022/7351061.

[27] M. Mohammadian, M. Babaei, M. Amin Jarrahi, and E. Anjomrouz, "Scheduling Nurse Shifts Using Goal Programming Based on Nurse Preferences: A Case Study in an Emergency Department," Int. J. Eng., vol. 32, no. 7, pp. 954–963, Jul. 2019, doi: 10.5829/ije.2019.32.07a.08.

[28] P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability," IEEE J Biomed Health Inform, vol. 17, no. 3, pp. 727–733, May 2013, doi: 10.1109/jbhi.2013.2244902.

[29] H. Liao et al., Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting: First International Workshop, MLMECH 2019, and 8th Joint International Workshop, CVII-STENT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings. Springer Nature, 2019. [Online].

[30] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE J Biomed Health Inform, vol. 18, no. 6, pp. 1750–1756, Nov. 2014, doi: 10.1109/JBHI.2014.2337752.

[31] S. L.Crawford, "Extensions to the CART algorithm," International Journal of Man-Machine Studies, vol. 31, no. 2, pp. 197-217, 1989.

[32] Zahra Boreiri,Alireza Norouzi Azad,Amin Ghodousian, "A Convolutional Neuro-Fuzzy Network Using Fuzzy Image Segmentation for Acute Leukemia Classification," 27th International Computer Conference, Computer Society of Iran (CSICC),IEEE, pp. 1-7, 2022.

[33] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic Methods to Extract New York Heart Association Classification from Clinical Notes," Proceedings, vol. 2017, pp. 1296–1299, Nov. 2017, doi: 10.1109/BIBM.2017.8217848.

[34] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," Int. J. Appl. Inf. Syst., vol. 3, no. 7, pp. 25–30, Aug. 2012, doi: 10.5120/ijais12-450593.

[35] "www.cdc.gov," 2020. [Online]. Available: https://www.cdc.gov/.

[36] Marvin Minsky, Seymour Papert, Perceptrons; an Introduction to Computational Geometry, United States: MIT Press, 1969

[37] Jagpreet Sidhu, Arvinder Kaur, Natural Language Processing. A Machine Learning Approach to Sense Tagged Words Using K-Nearest Neighbor, Germany: Bod Third Party Titles, 2018.

[38] C. Albon, Machine Learning with Python Cookbook,Practical Solutions from Preprocessing to Deep Learning, Japan: O'Reilly Media, 2018.

[39] Robert E. Schapire, Yoav Freund, Boosting,Foundations and Algorithms, United Kingdom: MIT Press, 2014.

[40] Y. L. Pavlov, Random Forests, Germany: VSP, 2000.

[41] C. Sheppard, Tree-based Machine Learning Algorithms,Decision Trees, Random Forests, and Boosting, United States: CreateSpace Independent Publishing Platform, 2017.

[42] N. V. Chawla,K. W. Bowyer,L. O. Hall,W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," The Journal of Artificial Intelligence Research (JAIR),16, pp. 321-357,2002.