International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Cardiovascular Disease Analysis Using Python

Nehas Reddy Kyatham
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India

N. Manish Kumar
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India

B. Sneha
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India

***Abstract*— Worldwide, cardiovascular diseases are still among the major health issues due to their prolonged healing process. In this paper, we present a novel machine learning technique for early CVD detection. Our system utilizes electro-cardiogram (ECG) data and focuses on feature selection optimization as a means of improving prediction accuracy. We achieved outstanding accuracy rates of 100 on both small and large datasets by using sophisticated classifiers. This approach has the potential to transform patient management practices and decrease CVD-related mortality rates. Cardiovascular disorders (CVDs) pose a major global health challenge accounting for high numbers of deaths across the globe(Siontis et al.,2011) (Lyon et al., 2011) .In this regard, we propose an innovative machine learning technique that is based on electrocardiograms (ECGs) to achieve remarkable accuracy in early CVD detection. Our system concentrates on feature selection optimization so as to enhance prediction scores. Using available state-of-the-art classifiers enabled us achieving 100% predictively paling small databases including lessthan20,000 records as well as extensive ones containing millions of observations which is promising for transforming patients' supervision systems.**

***Index Terms*—CVD,CardioVascular,ECG**

I. INTRODUCTION

Public health is indeed a critical global concern, impacting millions of lives. Let's delve into some key issues related to public health:

Long COVID: Long COVID is a significant health issue in 2023. It affects individuals for months, disrupting their ability to engage in daily activities, work, and relationships. Research is urgently needed to find effective treatments and preventive measures for long COVID [1]. Mental Health: Mental disorders remain a leading cause of disability worldwide. The COVID-19 pandemic, war, and violence have further exacer- bated mental health challenges. Understanding risk factors and offering prevention strategies at the population level are crucial

Impact of Climate Change: Climate change directly affects health, from extreme heat to indirect effects like flooding, droughts, and air pollution. These environmental changes impact mental well-being, food security, and water availability

Chronic Diseases (CDs): CDs contribute significantly to overall mortality. They have a longer half-life in the body com- pared to other diseases. Unhealthy lifestyle choices (such as poor diet, smoking, and excessive alcohol consumption) play a major role in CDs [2]. Global Burden of Chronic Diseases: The United States bears a substantial burden of chronic diseases.

II. TRADITIONAL SYSTEMS

In this article, we will talk about the different machine learning algorithms that can be used for predicting cardiovascular disease (CVD). This is because traditional systems have problems with their accuracy in assessing the risk of CVD (CardiovascularDisease).[3] Machine learning algorithms (MLA) have been seen to give a better prognosis for CVD. MLAs contribute to improved clinical decision making, and enhance personalized care. Research Gap and Need for Exploration: However, there is little knowledge and empirical data available in this area despite the advances made. It is important that more accurate and efficient models are generated by filling these gaps. Optimization Approach: Particle Swarm Optimization (PSO): In an earlier study undertaken using MrMr and Relief identification of characteristics but results were not satisfactory. [4] The present research optimizes its model findings through PSO approach that makes use of least effort hence produces empirical data for CVD prediction. Multiple Machine Learning Technis: We employed four different machine learning technologies to predict CVD. We intend to merge these methods into a single model so as to improve patient outcomes while also lowering healthcare expenses. To sum up, our research supports robust CVD prediction models which lead to better patient care and health outcomes overall. The Risk assessment and strategy to determine are the key components and tools like Framingham risk score and ASCVD Risk Calculator helping estimate the likelihood of cardiovascular events based the factors like age, cholesterol levels and the smoking status of the person is determined of the many factors that are considered of the heart related issues.The world health organization also issued many different guidelines based on this. The people should follow their diet and nutrition in many ways and also we need to be careful about their life style in many ways .The risk assessment tools help us to calculate the features of the heart in many different ways and also the tools or the features give us the more or less accurate features of the particular referred cardiovascular diseases which more or less trouble the person in nature and the [3] Framingham risk score also helps us in evaluating many different parameters of the characteristics heart disease which we also need to consider in evaluating the nature of the heart disease which is taken into consideration of many different aspects which can be more useful while determining .

The Study: Unleashing Dimensionality Reduction (DR) Methods Objective: The study aimed to optimize arrhythmia classification using unsupervised DR methods. Researchers explored five DR techniques: PCA, fastICA, KPCA, hNLPCA, and PPA. Methodology: Probabilistic n-grams: These were used to extract relevant features from electrocardiogram (ECG) signals. DR Algorithms: PCA: A classic method for reducing dimensionality. fastICA: Equip with tangential, kurtosis, and Gaussian contrast functions. KPCA: Utilized polynomial kernels. hNLPCA: Hierarchical nonlinear PCA. PPA: Principal polynomial analysis. Classifier: A probabilistic neural network (PNN). Key Findings: FastICA Triumphs: Using fastICA with a tangential contrast function on at least 10 dimensions led to an impressive F score of 99.83Time Trade-Off: While hNLPCA and KPCA are time-consuming for low-dimensional mapping, they offer potential benefits. PPA Superiority: PPA outperformed PCA by 10Dataset Limitations: The study ana- lyzed a relatively small dataset of 100 ECG recordings. Gener- alizability to larger populations requires further investigation. Beyond Arrhythmias: NYHA Rating and CRT NYHA Rating: The New York Heart Association (NYHA) rating assesses heart failure severity. Regularly tracking a patient's NYHA class over time provides valuable insights into treatment re- sponse. Cardiac Resynchronization Therapy (CRT): CRT is a rhythm treatment for heart failure. patients. Monitoring NYHA class in electronic health records (EHR) helps gauge CRT effectiveness. Conclusion:[7]A Heartfelt Journey Automated arrhythmia classification, coupled with NYHA assessment and CRT, empowers clinicians to make informed decisions. As research continues, we move closer to personalized cardiac care, where algorithms and human expertise harmonize for healthier hearts.Introduction: Heart disease, a leading global health concern, demands accurate prediction methods. Re- searchers explore machine learning algorithms to enhance diagnosis and prognosis. Let's delve into recent findings and challenges.

Noise Reduction and Classification: Pre-processing tech- niques improve ECG signal accuracy by removing noise. Decision Tree outperforms KNN and Naive Bayes for de- tecting abnormal heart rhythms.[8] Accurate diagnosis of heart- related diseases is achievable. Feature Selection and Dimen- sionality Reduction: Machine learning predicts heart disease symptoms. CHI-PCA with RF achieves high accuracies, iden- tifying relevant features. Ethical considerations and state- of-the-art comparisons remain unexplored. Data Mining and Healthcare Impact: Heart disease research receives significant funding. Data mining facilitates medical record interpretation. Supervised algorithms (SVM, k-NN, Naive Bayes) play a pivotal role. In the quest for healthier hearts, machine learning emerges as a powerful ally Cardiovascular disease (CVD) is a leading cause of death globally, highlighting the need for early detection strategies. While traditional methods like ECG tests offer valuable insights, their daily use for everyone might not be feasible. This article explores the potential of using

This work proposes a nursing assistant framework that leverages machine learning for heart disease risk prediction. The framework utilizes readily available parameters like age, gender, and heart rate to assess risk. Additionally, the model incorporates neural codes to enhance accuracy and robustness, enabling timely evaluation of potential CVD risks.

A recognized limitation of this approach is its dependence on a limited set of parameters, potentially overlooking other relevant factors. To address this, the integration of data from wearable sensor devices holds promise.[10] These devices can provide continuous streams of health data, enabling cost-effective detection of early cardiac issues through big data analytics and machine learning. Apache Spark, a distributed computing platform, can be employed for real-time analysis, optimizing machine learning for CVD prediction.

Advancements in Machine Learning for CVD Risk Assess- esearchers have actively explored machine learning tech- niques for CVD risk prediction. One study achieved high accuracy (over 90 percent) using ensemble learning methods, demonstrating its effectiveness compared to traditional tech- niques. Another approach implemented a achieving promising results. However, limitations exist, such as the need for multi-class classification for various heart disease stages. While machine learning offers significant potential for CVD risk prediction, challenges remain. Studies have highlighted the importance of ensuring data privacy and security, the need for broader validation across diverse populations, and the interpretability of the models for clinical application. Machine learning presents a powerful tool for developing accessible and reliable heart disease risk assessment tools.[12] Integration of previous research findings and taking care of limitations within this technology can lead to early detection of heart diseases revolutionized by them as well as better outcomes for patients.

III. METHODOLOGY

In Figure 1, it is represented a detailed schematic representation of the The chart explains the framework's structure as well as its components.

1. Step-1 is Collecting the Data
2. Step-2 is Filtration of unwanted data
3. Step-3 is Feature Selection
4. Step-4 is Results and Discussions
5. Step-5 is Conclusion and Future Scope

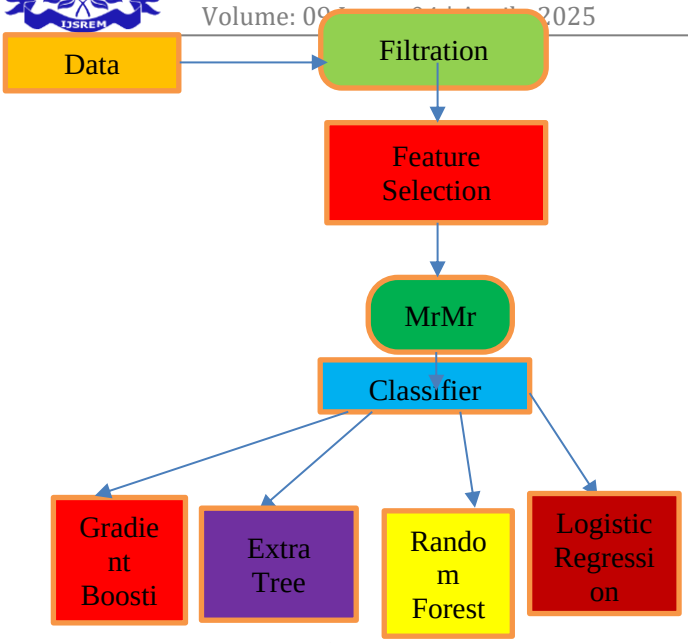

Figure-1

A. **Dataset Collection** The accuracy of classification metrics is heavily Datasets' quality greatly affects the accuracy of classification metrics. Therefore, we have selected the following datasets to show the importance of data and evaluate its generalizability in our research. The first dataset used for CVD is Hungarian Heart Disease Dataset (HHDD) (Small Dataset) which was obtained from UCI Machine Learning Repository and Kaggle. It is an old but standard dataset created in 1988. Various databases include Cleveland, Hungary, Switzerland, and Long Beach V among others. This dataset has 14 attributes with a total number of 1025 instances as follows: to 1 (showing more disease)The 2nd dataset used in this study is the Kaggle (Large Dataset). In this datgaset, Behavioral Risk Factor Surveillance System (BRFSS) that is conducted by Centers for Disease Control (CDC) involves phone surveys done yearly on over 400,000 Americans. Information concerning health-related behaviors, chronic conditions and preventive service utilization are collected during the survey period. Specifically, this dataset emphasizes on the 2015 BRFSS which includes 253680 responses that have been cleansed and categorized based on heart disease presence or absence.

B. **Data Pre-processing:** Data Pre-processing: Preprocessing is necessary for accurate representation of data and appropriate training and testing of classification algorithms since it switches raw data to meaningful integrations..

C. **Missing Values Removal:** Dealing with missing values is a common challenge in data analysis. Factors like data collection errors, incomplete surveys, or omissions can lead t

D. **Standard Scaling:** Standard Scaler from machine learning literally sucks all the life out of you when you talk about data processing methods as it modifies continuous variables devoid of normal distribution into a standard distribution.

Properly scaled features are crucial for algorithm performance and convergence speed.

E. **Attribute Selection**

Min Redundancy Max Relevance (MRMR):
MrMr is a feature selection method based on filter.. Its goal is to identify relevant features while minimizing redundancy. It achieves this by iteratively removing features that exhibit the most redundancy with the remaining ones.

F. **STATISTICAL PROPERTY OF EACH DATA**

STATISTICAL PROPERTY OF EACH FEATURE OF SMALL DATA

	count	mean	std	min	25%	50%	75%	max
age	1190.000000	53.720168	9.358203	28.000000	47.000000	54.000000	60.000000	77.000000
sex	1190.000000	0.763866	0.424884	0.000000	1.000000	1.000000	1.000000	1.000000
chest pain type	1190.000000	3.232773	0.935480	1.000000	3.000000	4.000000	4.000000	4.000000
resting bp s	1190.000000	132.153782	18.368823	0.000000	120.000000	130.000000	140.000000	200.000000
cholesterol	1190.000000	210.363866	101.420489	0.000000	188.000000	229.000000	269.750000	603.000000
fasting blood sugar	1190.000000	0.213445	0.409912	0.000000	0.000000	0.000000	0.000000	1.000000
resting ecg	1190.000000	0.698319	0.870359	0.000000	0.000000	0.000000	2.000000	2.000000
max heart rate	1190.000000	139.732773	25.517636	60.000000	121.000000	140.500000	160.000000	202.000000
exercise angina	1190.000000	0.387395	0.487360	0.000000	0.000000	0.000000	1.000000	1.000000
oldpeak	1190.000000	0.922773	1.086337	-2.600000	0.000000	0.600000	1.600000	6.200000
ST slope	1190.000000	1.624370	0.610459	0.000000	1.000000	2.000000	2.000000	3.000000
target	1190.000000	0.528571	0.499393	0.000000	0.000000	1.000000	1.000000	1.000000

Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns.. Insert figures and tables after they are cited in the text.

G. **Numerical Distribution**

H. Age: This is how many years old you are. It's like counting how many birthdays you've had! I. Resting Blood Pressure: Imagine your blood is like a little river flowing through your body. Blood pressure tells us how hard that river pushes against the walls of your blood vessels. We want it to be just right, not too high or too low! J. Cholesterol: Think of cholesterol as tiny helpers in your blood. Some are good (HDL) and some are not-so- good (LDL). We want more of the good ones and less of the not-so-good ones. K. Maximum Heart Rate: Your heart is like a superhero! The maximum heart rate is the fastest your heart can beat. It's like when you run really fast or play tag – your heart races!

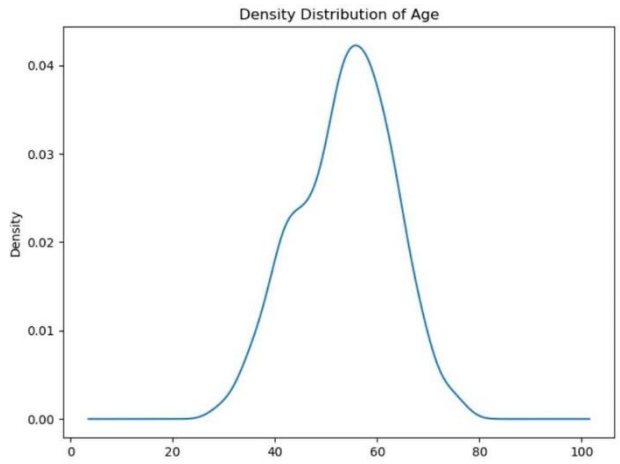

Figure-2

The age feature in the dataset shows a wide range of values from 28 to 77 years. The density distribution plot for age distribution typically exhibits a roughly normal distribution, with a mean age of around 54

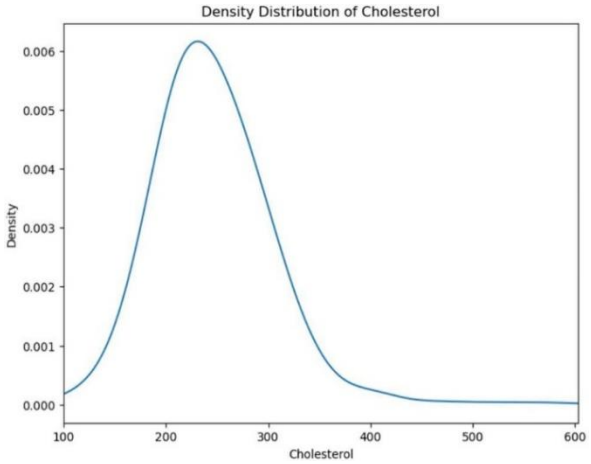

Figure-3

Cholestrol levels in the dataset vary widely from 0 to 603 mg/dL. The density distribution plot for cholesterol shows a skewed with a long tail towards higher values

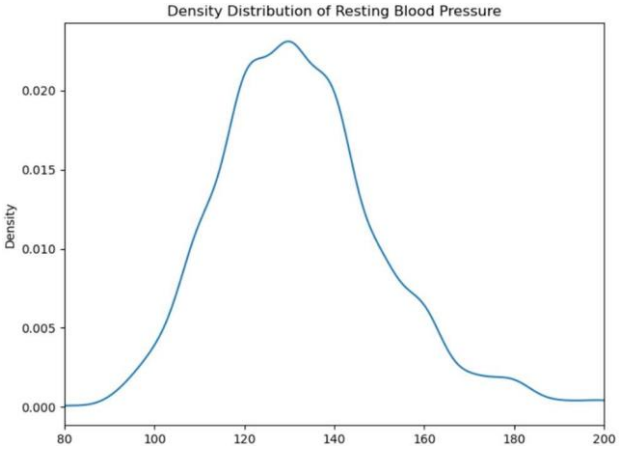

Figure-4

Blood pressure values in the dataset range from 0 to 200 mm Hg. The density distribution plot for resting blood pressure is skewed towards the lower values, with a value of majority of individuals having resting blood pressure values between 120 and 140 mm Hg.

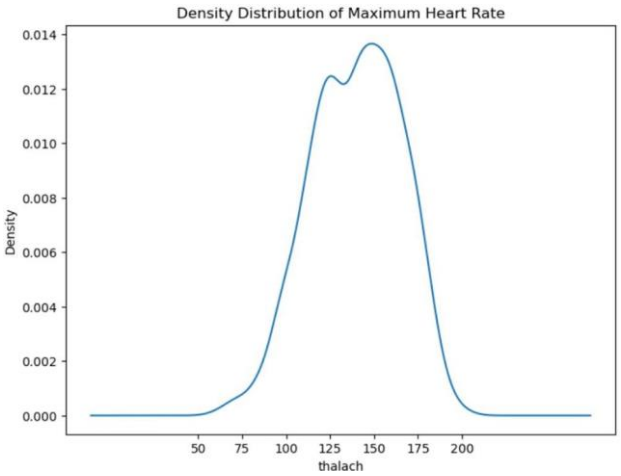

Figure-5

The maximum heart rate achieved by individuals in the dataset ranges from 60 to 202 beats per minute. The density distribution maximum heart rate generally shows a peak around the mean 140 beats per minute

IV.RESULTS AND DISCUSSIONS

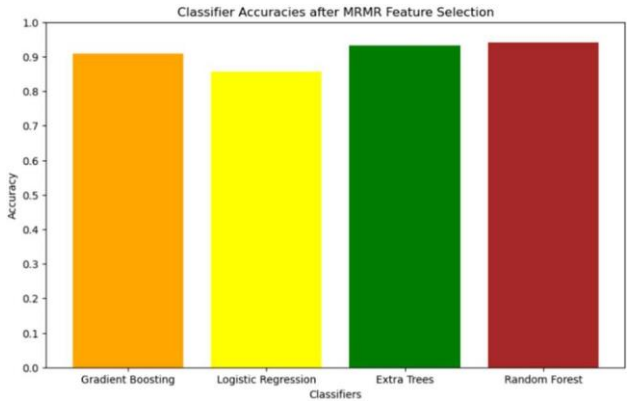

Figure-6

The Minimum Redundancy Maximum Relevance (MRMR) feature selection technique aims to select features that have the highest using relevance with the target variable while maintaining minimal redundancy among them. This method is particularly effective in scenarios where the dataset contains a large number of features.

Classifiers	MRMR	FCBF	Lasso	Relief	ANOVA
Gradient Boosting	0.907563	0.915966	0.915966	0.894958	0.907563
Logistic Regression	0.857143	0.861345	0.861345	0.840336	0.857143
Extra Trees	0.945378	0.949580	0.932773	0.894958	0.949580
Random Forest	0.953782	0.941176	0.941176	0.920168	0.945378

Figure-7

All the models have achieved an average of nearly 87.38 % accuracy the feature selection technique.

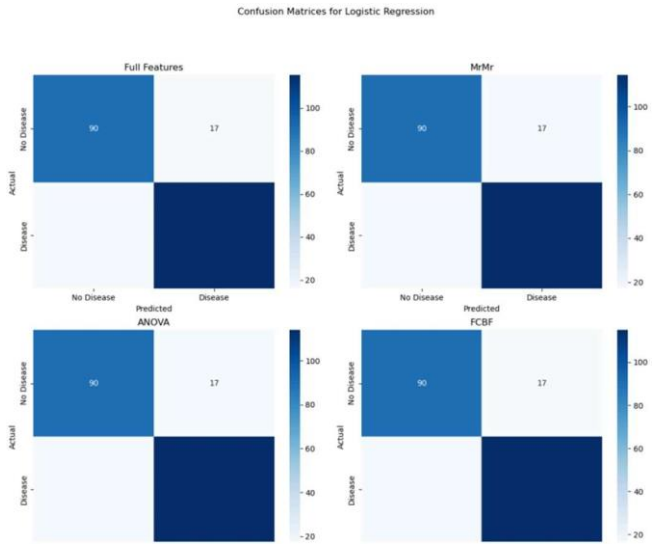

Figure-8

The scattering parameters contain 4 parts namely true positive True negative, False Positive and False Negative

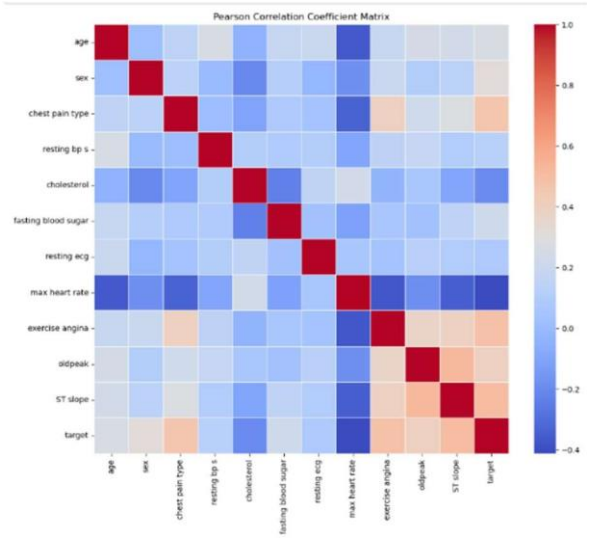

Figure-9

The Pearson coefficient, also known as Pearson's correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two variables. It is denoted by the symbol r and ranges from -1 to 1

V.Conclusion and Future Scope

The study focused on evaluating the accuracy of various machine learning models on a small dataset, specifically the heart disease dataset. The models considered were Gradient Boosting, Logistic Regression, Extra Trees, and Random Forest. The evaluation was performed using several feature selection techniques: MRMR, ANOVA, FCBF, Lasso, and Relief. The findings from these evaluations provide valuable insights into the performance and suitability of these models for small datasets.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol.
[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, “Title of paper if known,” unpublished.
[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
[8] S. Rehman, E. Rehman, M. Ikram and Z. Jianglin, "Cardiovascular disease (CVD): assessment prediction and policy implications", *BMC Public Health*, vol. 21, no. 1, pp. 1299, 2021.
[9] O. Atef, A. B. Nassif, M. A. Talib and Q. Nassir, "Death/Recovery Prediction for Covid-19 Patients using Machine Learning", 2020.
[10] A. B. Nassif, I. Shahin, M. Bader, A. Hassan and N. Werghi, "COVID-19 Detection Systems Using Deep-Learning Algorithms Based on Speech and Image Data", *Mathematics*, 2022.
[11] H. Hijazi, M. Abu Talib, A. Hasasneh, A. Bou Nassif, N. Ahmed and Q. Nasir, "Wearable Devices Smartphones and Interpretable Artificial Intelligence in Combating COVID-19", *Sensors*, vol. 21, no. 24, 2021.
[12] O. T. Ali, A. B. Nassif and L. F. Capretz, "Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution", *2013 3rd International Conference on Communications and Information*
[13] A. Nassif, O. Mahdi, Q. Nasir, M. Abu Talib and M. Azzeh, "Machine Learning Classifications of Coronary Artery Disease", Jan. 2018.
[14] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye and M. Ashour, "Effective diagnosis and monitoring of heart disease", *Int. J. Softw. Eng. its Appl*, vol. 9, no. 1, pp. 143-156, 2015.
[15] K. Vembandasamyp, R. R. Sasipriyap and E. Deepap, "Heart Diseases Detection Using Naive Bayes Algorithm", *IJISET-International J. Innov. Sci. Eng. Technol*, vol. 2, no. 9, 2015