AI Driven Approach to Reducing Cardiovascular Risk

MARISETTI LAKSHMI PRASANNA[1], K V DURGA DEVI[2]
#1. M.Tech Scholar in Department of Artificial Intelligence,
#2. Assistant Professor, Department of Artificial Intelligence, Kakinada Institute Of Engineering & Technology or Women, AP, India.

ABSTRACT
Objectives: *The latest statistics of the World Health Organization anticipated that cardiovascular diseases including Coronary Heart Disease, Heart attack, vascular disease as the biggest pandemic to the world due to which one-third of the world population would die. With the emerging AI trends, applying an optimal machine learning model to target early detection and accurate prediction of heart disease is indispensable to bring down the mortality rates and to treat cardiac patients with the best clinical decision support. This stems from the motivation of this paper. This paper presents a comprehensive survey on heart disease prediction models derived and validated out of popular heart disease datasets like the Cleveland dataset, Z-Alizadeh Sani dataset.* ***Methods:*** *This survey was performed using the articles extricated from the Google Scholar, Scopus, Web of Science, Research Gate, and PubMed search engines between 2005 to 2020. The main keywords for the search were Heart Disease, Prediction, Coronary disease, Healthcare, Heart datasets, and Machine Learning. Results: This review explores the shortcomings of various approaches used for the prediction of heart diseases. It outlines the pros and cons of different research methodologies along with the validation parameters of each reviewed publication.* ***Conclusion:*** *Machine intelligence can serve as a genuine alternative diagnostic method for prediction, which will, in turn, keep the patients well aware of their illness state. Despite the researcher's efforts, still uncertainty exists towards the standardization of prediction models which demands further exploration of optimal prediction models.*

Keywords: *Heart diseases, Machine learning, Deep learning, Health care, Heart disease dataset*

I. INTRODUCTION

Heart Disease/Disorders (HD) have been recognized as one of the convoluted and fatal human illnesses in the world. Due to this disease, the heart functions abnormally leading to blocked blood vessels and get affected by angina, heart attack, and stroke. The most common types of heart diseases are Coronary Vascular Disease (CVD), Coronary Artery Disease (CAD), Congestive Heart Failure (CHF), and Abnormal Heart Rhythms. There are many challenges in predicting such HD at the early stages due to the involvement of several conventional risk factors like age, sex, hypertension, high cholesterol, abnormal pulse, and many other factors [1]. Despite wide diversity in the existence of cardiovascular risk factors across different sectors of society, CVD has been noticed to be one of the major causes of death all over India including economically backward states and rural areas. The global statistics also showed that the premature mortality in terms of years of life lost because of CVD climbs to 37 million (2010) from 23.2 million (1990) with an incremental rise of 59 % every year, which serves as the prime motivation of this paper.

The need for heart disease diagnosis has compelled towards invention few invasive clinical techniques like angiogram, which in spite of being expensive also induces some side effects for the diagnosed patients. This has motivated several researchers to use data mining techniques to diagnose CVD safely.

Machine Intelligence is a type of intelligence exhibited by machines to interconnect with the physical world [2].

Machine learning and deep learning technologies are two subsets of AI, which are likely to be used as the model to predict and ascertain the data. Both these technologies are very powerful and worthy for medical data analytics. Application of different types of machine intelligence paradigms is an ideal approach for heart disease diagnosis but as well serves as an aid for prediction, illness monitoring, and its other related clinical management aspects [3,4].

The related works of machine/deep learning in the medical field related to heart disease predictions have been explored elaborately in forthcoming sections and the generalized framework opted by most of the researchers for the prediction of heart disorders is shown in Figure 1. A prelude on the heart disease datasets commonly used by the researchers is presented in the subsequent section.

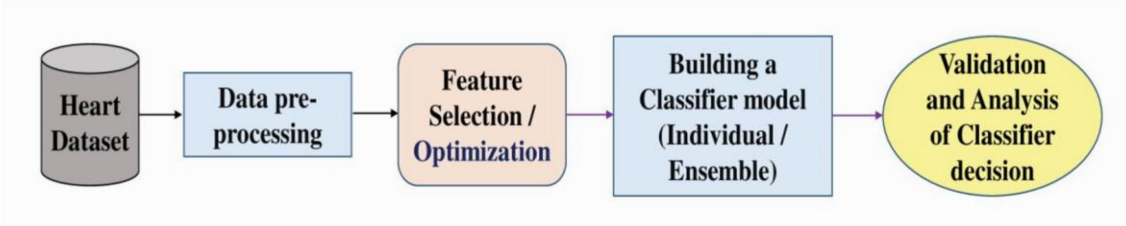
Heart Dataset
Data pre-processing
Feature Selection / Optimization
Building a Classifier model (Individual / Ensemble)
Validation and Analysis of Classifier decision

Figure 1 Generalized heart disease prediction framework

This article provides the benefits and shortcomings of the reviewed publications in the results section and highlights the salient points in the discussion section.

HEART DISEASE DATASETS

This section provides an overview of datasets commonly used in the reviewed publications.
The most popular dataset used by the researchers is the Cleveland heart disease dataset obtained from the online repository of the University of California, Irvine (UCI) for machine learning. It is comprised of 303 samples with 6 samples having missing values. The data, in its original form, have 76 features but all the published work is likely to refer to 13 features out of them and the other feature outlines the effect of the disease. The salient features with their valid ranges are presented in Table 1.

Table 1 Cleveland dataset description

S. No.	Attribute	Description	Range
1	Age	Age of the individual	29-77
2	Sex	Sex	M, F
3	CP	Chest Pain type	1-typical angina
			2-atypical angina
			3-Non-Anginal Pain
			4-Asymptomatic
4	restbp	Resting Blood Pressure	94-200
5	serchol	Serum Cholestoral in mg/dl	126-564
6	fbs	Fasting blood sugar > 120	Yes, No
7	restecg	Resting Electrocardiographic	0, 1, 2
8	mhr	Maximum Heart rate achieved	71-202
9	exang	Exercise Induced Angina	Yes, No
10	oldpeak	ST depression Induced by Exercise relative to Rest	0-6.2
11	slope	Slope of the Peak Exercise ST Segment	1, 2, 3
12	vca	Number of Major Vessels colored by Fluoroscopy	0, 1, 2, 3
13	thal	Thallium Scan	3-Normal
			6-Fixed Defect
			7-Reversible Defect
14	num	Diagnosis of heart disease	0: <50% diameter narrowing
			1: >50% diameter narrowing

Another most prevalent dataset used by the researchers for the prediction process is the Z-Alizadeh Sani dataset that includes 303 patients' data with 55 input variables and a class label variable of each patient. The class label variable is comprised of four groups i.e., normal, LAD, LCX, and RCA which all come into the category of coronary heart disease. This dataset was mainly assembled for the diagnosis of CAD. The features, along with their valid ranges are introduced in Table 2.

Table 2 Z-Alizadeh Sani dataset description

Feature Type	Feature Name	Range
Demographic	Age	30-86
	Weight	48-120
	Sex	Male, Female
	BMI (Body Mass Index Kg/m^2)	18-41
	DM (Diabetes Mellitus)	Yes, No
	HTN (Hypertension)	Yes, No
	Current Smoker	Yes, No
	Ex-Smoker	Yes, No
	FH (Family History)	Yes, No
	Obesity	Yes if MBI>25, No otherwise
	CRF (Chronic Renal Failure)	Yes, No
	CVA (Cerebrovascular Accident)	Yes, No
	Airway Disease	Yes, No
	Thyroid Disease	Yes, No
	CHF (Congestive Heart Failure)	Yes, No
	DLP (Dyslipidemia)	Yes, No
Symptom and Examination	BP (Blood Pressure: mmHg)	90-190
	PR (Pulse rate: ppm)	50-110
	Edema	Yes, No
	Weak Peripheral Pulse	Yes, No
	Lung Rales	Yes, No
	Systolic Manner	Yes, No
	Diastolic Manner	Yes, No
	Typical Chest Pain	Yes, No
	Dyspnea	Yes, No
	Function Class	1, 2, 3, 4
	Atypical	Yes, No
	Nonanginal Chest Pain	Yes, No
	Exertional Chest Pain	Yes, No
	Low Th Ang (Threshold Angina)	Yes, No
ECG	Rhythm	Sin, AF

	Q wave	Yes, No
	ST-Elevation	Yes, No
	ST Depression	Yes, No
	T inversion	Yes, No
	LVH (Left Ventricular Hypertrophy)	Yes, No
	Poor R Progression	Yes, No
Laboratory and Echo	FBS (Fasting Blood Sugar: mg/dl)	62-400
	Cr (Creatine: mg/dl)	0.5-2.2
	TG (Triglyceride: mg/dl)	37-1050
	LDL (Low Density Lipoprotein: mg/dl)	18-232
	HDL (High Density Lipoprotein: mg/dl)	15-111
	BUN (Blood Urea Nitrogen: mg/dl)	6-52
	ESR (Erythrocyte Sedimentation rate: mm/h)	1-90
	HB (Haemoglobin: g/dl)	8.9-17.6
	K (Potassium: mEq/lit)	3.0-6.6
	Na (Sodium: mEq/lit)	128-156
	WBC (White Blood Cells: cells/ml)	3700-18000
	Lymph (Lymphocyte) (%)	Jul-60
	Neut (Neutrophil) (%)	32-89
	PLT (Platelet: 1000/ml)	25-742
	EF (Ejection Fraction) (%)	15-60
	Region with RWMA (Regional Wall Motion Abnormality)	0, 1, 2, 3, 4
	VHD (Valvular Heart Disease)	Normal, Mild, Moderate, Severe

The other datasets that are used by the researchers in the prediction process are StatLog Heart, Hungarian, Long Beach VA, and Kaggle Framingham dataset. StatLog dataset consists of 270 samples and each sample has 13 features similar to Cleveland as presented in Table 1.

The other two datasets of Hungarian and Long Beach VA datasets are obtained from the UCI repository where each dataset consists of 274 samples with each of 14 features like the Cleveland dataset presented in Table 1. In the Kaggle Framingham dataset, a large amount of data is available with samples of 4240 patients comprising of 16 features that incorporate behavioral, demographic, and medical risk factors.

II. RESULTS

In recent years, there have been ample investigations by several researchers on heart disease predictions using the above-mentioned available datasets.

In the year of 1979, GA Diamond, JS Forrester integrated different results obtained from tests like stress electrocardiography, cardiokymography, thallium scintigraphy, and cardiac fluoroscopy into a diagnostic conclusion about the probability of acquiring disease in a given patient using Bayes' Theorem [5]. Later the heart disease approaches have taken a new dimension towards estimation of the CHD using risk factor categories with the help of regression equations and logistic methods by WF Wilson, et al. [6].

In the later stages, different machine learning and deep learning algorithms are developed by several researchers to predict cardiovascular disease on the datasets available in the UCI repository.

In this paper, some of the publications related to heart disease predictions have been reviewed.

The comparative analysis of several reviewed works related to heart disease prediction is presented in Table 3.

Table 3 Comparison of various heart disease prediction approaches

Ref.	Year	Classifiers/ methods	Dataset Used	No. of Selected attributes	Inferences	
					Benefits	Drawbacks
[7]	2007	Feed Forward back propagation network	Not Specific	13	Unlabelled data fed for obtaining the classification accurately. 100% accuracy achieved.	Less size of data (78 records) No performance metric is evaluated. Human involvement testing is preferred.
[8]	2007	TAN, STAN, C4.5, CMAR and SVM	Not Specific	8	SVM showed the best accuracy of 90.9% among all.	Less size of data (193 records) The accuracy of each classifier is varied for three different recumbent positions.
[9]	2009	Neural networks using LM, SCG, and CGP algorithms	Cleveland	13	Classification accuracy is 89.01%. Specificity is 95.91%.	No other performance metric is evaluated to standardize the results Spare Complexity.
[10]	2010	NB, DL, KNN	Kaggle Framingham	14	Huge dataset (4240 instances) Naïve Bayes performed well compared with others and achieved an accuracy of 52.33%.	Obtained accuracy is very less (52.33%) compared to all. No other performance metric is evaluated.
[11]	2013	Backpropagation network	Cleveland	13	Obtained accuracy is 92% at the 10th run time of the algorithm with different seed numbers.	Less size of training (116 records) and testing data (50 records) No feature extraction process.
[12]	2014	NB, DT-GI, SVM	Cleveland	13	The accuracy of the majority voting based ensemble is 81.82% Specificity is 92.86%	The type of feature selection and no. of attributes selected for the ensemble process are not mentioned.
[13]	2017	Multiple Feature Selection with an ensemble approach	Z-Alizadeh Sani	34	Obtained accuracy is 93.70 ± 0.48 %. F1-score (95.53%) and Recall (97.63%) provide the best results.	Processing time is high. No significance test is considered. Sparse complexity.
[14]	2017	Bagged Tree, Adaboost, and RF	Statlog	7	Feature selection technique is utilized Bagged tree with PSO provides a classification accuracy of 100%. Recall and specificity are 100%	The size of the dataset is less. Not compared with other datasets for standardization of the obtained result.
[15]	2017	An adaptive weighted fuzzy system using GA+MDMS-PSO	Cleveland, Hungarian, and Switzerland	7	Three different statistical methods are used to identify the risk level factors Experiments were carried out on different datasets and achieved accuracies of 92.31% for Cleveland, 95.56 % for Hungary, 89.47% for Switzerland, 91.8% for Long Beach, and 92.68% for Heart datasets.	The suggested model gives preferable results for one statistical method. The generalization of the system is not guaranteed as fewer performance metrics are evaluated. No significance tests are performed.
[16]	2018	LR, KNN, ANN, SVM, NB, DT	Cleveland	6	Three different feature selection algorithms are used and compared with the classifiers. The best accuracy is 89% with LR and Relief algorithm.	Despite using several metrics the best algorithm is varied for all metrics. A single dataset is used for the entire process and no comparisons are made with other datasets.
[17]	2018	KNN, RF, SVM, NB & ANN	Statlog	7	FCBF feature selection method is used besides two optimization approaches namely PSO and ACO. Accomplished with an accuracy of 99.65% using KNN	Each algorithm worked worse in some situations. A single dataset is used for the entire process and no comparisons are made with other datasets.
[18]	2019	NB, GLM, LR, DL, DT, RF, GBT & SVM	Cleveland	13 (8 Subsets)	The hybrid algorithm is implemented with RF and LM. Achieved the best accuracy of 88.47%	A single dataset is used for the entire process and no comparisons are made with other datasets. The age factor is excluded from the modeling.
[19]	2019	NuSVM, LinSVM and SVC	Z-Alizadeh Sani	29	Two-level optimization is preferred using GA and PSO Results are given that the highest accuracy is obtained for nuSVM (93.08%).	No significance test is conducted for the standardization of the proposed approach. The exactness is less on the same dataset refer to [10]
[20]	2019	NB, BN, RF and MLP	Statlog	11 (6 Subsets)	The maximum increase of 7% accuracy is achieved by the majority vote ensemble.	The total complexity is not determined. The age factor is excluded from the model.

					The computational time for ensemble techniques is determined. Experiments are done and achieved an accuracy of 85.48% by an ensemble of BN, NB, RF, and MP.	No standardization has been proposed by comparing different datasets in the approach.
[21]	2019	χ^2-model+Deep Neural Network	Cleveland	11	The system was evaluated using 6 different performance metrics Comparisons were made between conventional neural networks (ANN, DNN) and proposed neural networks (χ^2-ANN and χ^2-DNN Under-fitting and overfitting problems are resolved. Achieved a testing accuracy of 93.33%	A single dataset with small sample size is used to test the system. The time complexity is not determined. The search strategy is used for the optimal width selection for hidden layers in ANN and DNN.
[22]	2019	Random Search+Random Forest	Cleveland	7	Reduces the time complexity as the number of features is reduced. Achieved a testing accuracy of 93.33% The overfitting problem is resolved	A single dataset with small sample size is used to test the system. Specific processing time is not mentioned in the approach.
[23]	2019	χ^2-model+Gaussian NB	Cleveland	9	Six evaluation metrics are used for the Cleveland dataset. Achieved a testing accuracy of 93.33%	The age factor is excluded from the analysis. Time complexity is not determined.
[24]	2020	BiLSTM – CRF	Cleveland	-	Analyzed the data in both guiding ways and provide a linear relationship between attributes. Achieved a good classification accuracy of 90.04% for the Cleveland dataset. The proposed method is tested on 4 different datasets.	No. of attributes selected for the prediction is not clear. Average accuracy results are preferred over individual accuracies of different datasets. No significance test is calculated.

III. CONCLUSION

Machine intelligence can serve as a genuine alternative diagnostic method for prediction, which will, in turn, keep the patients well aware of their illness state.

This article presents a comprehensive study of heart prediction systems based on machine learning, ensemble, and deep learning approaches. From the reviewed literature, it is obvious that the Cleveland heart disease dataset that contains only 303 instances with 14 features is mostly used. This is mainly because of the tiny and restricted sample size. Any study that uses other data sources also concentrated on a single dataset with a limited number of features. Consequently, high accuracies obtained in the prediction models with the removal of irrelevant features or removal of highly correlated factors or by using feature selection/ optimization techniques cannot be generalized, which is a major shortcoming.

Despite the researcher's efforts, still uncertainty exists towards the standardization of prediction models. To get a more generalized classification and prediction accuracy, other multiple heart disease datasets from different sources with more features should be considered. An efficient predictive framework model which eliminates most of the shortcomings reported in this paper is the cardinal intent of our future research. Furthermore, real-time data should be analyzed on the working learning model to get it standardized and ensure its reliability with the clinical correlation and validation.

REFERENCES

[1]. Kusuma, S., and J. Divya Udayan. "Machine learning and deep learning methods in Heart Disease (HD) research." International Journal of Pure and Applied Mathematics, Vol. 119, No. 18, 2018, pp. 1483-1496.

[2]. Krittanawong, Chayakrit, et al. "Artificial intelligence in precision cardiovascular medicine." Journal of the American College of Cardiology, Vol. 69, No. 21, 2017, pp. 2657-64.

[3]. "Anticipating artificial intelligence." Nature, Vol. 532, No. 7600, 2016, p. 413.

[4]. Remeseiro, Beatriz, and Veronica Bolon-Canedo. "A review of feature selection methods in medical applications." Computers in Biology and Medicine, Vol. 112, 2019, p. 103375.

[5]. Diamond, George A., and James S. Forrester. "Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease." New England Journal of Medicine, Vol. 300, No. 24, 1979, pp. 1350-58.

[6]. Wilson, Peter WF, et al. "Prediction of coronary heart disease using risk factor categories." Circulation, Vol. 97, No. 18, 1998, pp. 1837-47.

[7]. Guru, Niti, and Anil Dahiya. "NavinRajpal "Decision support system for heart diseases prediction using neural networks" Delhi Business Review, Vol. 8, No. 1, 2007, pp. 1-6.

[8]. Lee, Heon Gyu, Ki Yong Noh, and Keun Ho Ryu. "Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2007.

[9]. Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert Systems with Applications, Vol. 36, No. 4, 2009, pp. 7675-80.

[10]. Rajkumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using datamining algorithm." Global Journal of Computer Science and Technology, Vol. 10, No. 10, 2010, pp. 38-43.

[11]. Al-Milli, Nabeel. "Backpropagation neural network for prediction of heart disease." Journal of Theoretical and Applied Information Technology, Vol. 56, No. 1, 2013, pp. 131-35.

[12]. Bashir, Saba, Usman Qamar, and M. Younus Javed. "An ensemble based decision support framework for intelligent heart disease diagnosis." International Conference on Information Society (i-Society 2014), IEEE, 2014.

[13]. Qin, Cai-Jie, Qiang Guan, and Xin-Pei Wang. "Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection." Biomedical Engineering: Applications, Basis and Communications, Vol. 29, No. 06, 2017, p. 1750043.

[14]. Yekkala, Indu, Sunanda Dixit, and M. A. Jabbar. "Prediction of heart disease using ensemble learning and Particle Swarm Optimization." 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), IEEE, 2017.

[15]. Paul, Animesh Kumar, et al. "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease." Applied Intelligence, Vol. 48, No. 7, 2018, pp. 1739-56.

[16]. Haq, Amin Ul, et al. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems, Vol. 2018, 2018.

[17]. Khourdifi, Youness, and Mohamed Bahaj. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." International Journal of Intelligent Engineering and Systems, Vol. 12, No. 1, 2019, pp. 242-52.

[18]. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE Access, Vol. 7, 2019, pp. 81542-54.

[19]. Abdar, Moloud, et al. "A new machine learning technique for an accurate diagnosis of coronary artery disease." Computer Methods and Programs in Biomedicine, Vol. 179, 2019, p. 104992.

[20]. Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked, Vol. 16, 2019, p. 100203.

[21]. Ali, Liaqat, et al. "An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network." IEEE Access, Vol. 7, 2019, pp. 34938-45.

[22]. Javeed, Ashir, et al. "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection." IEEE Access, Vol. 7, 2019, pp. 180235-43.

[23]. Ali, Liaqat, et al. "A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive bayes." Computational and Mathematical Methods in Medicine, Vol. 2019, 2019.

[24]. Manur, Manohar, Alok Kumar Pani, and Pankaj Kumar. "A prediction technique for heart disease based on long short term memory recurrent neural network." International Journal of Intelligent Engineering and Systems, Vol. 13, No. 2, 2020, pp. 31-39.

[25]. Garate-Escamila, Anna Karen, Amir Hajjam El Hassani, and Emmanuel Andres. "Classification models for heart disease prediction using feature selection and PCA." Informatics in Medicine Unlocked, Vol. 19, 2020, p. 100330.

[26]. Johnson, Kipp W., et al. "Artificial intelligence in cardiology." Journal of the American College of Cardiology, Vol. 71, No. 23, 2018, pp. 2668-79.

[27]. American Health Association. https://www.heart.org/

[28]. Prabhakaran, Dorairaj, Panniyammakal Jeemon, and Ambuj Roy. "Cardiovascular diseases in India: Current epidemiology and future directions." Circulation, Vol. 133, No. 16, 2016, pp. 1605-20.

[29]. Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." JAMA, Vol. 316, No. 22, 2016, pp. 2402-10.