2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)

Computational Approaches for Prediction of Cardiovascular Risks along with API Feed

Pratham Agrawal
SCOPE
Vellore Institute of Technology
Vellore, India
prathamagrawal1205@gmail.com

Vaibhav Bhartia
SCOPE
Vellore Institute of Technology
Vellore, India
bhartiavaibhav@gmail.com

Tusar Kanti Mishra
SCOPE
Vellore Institute of Technology
Vellore, India
tusar.k.mishra@gmail.com

***Abstract*—Cardiovascular diseases (CVD) refer to a variety of ailments that affect the coronary heart and are the main reason for mortality in the latest decades. Predicting the possibility of cardiac contamination is a difficult task, automating it can help treat the ailment early and efficiently. The prognosis of coronary heart disease is based entirely on previous clinical records and has been deemed unreliable due to several factors. Non-invasive strategies together with machine learning methods can be used to classify those with cardiac problems efficiently. In this paper, Internet of Things (IoT) has been utilized interfaced input parameters through suitable models for low latent prediction of heart ailments. Using the 2016 Cleveland Heart Diseases dataset, models have been proposed, which predict the disease with good accuracy. Using data engineering along with machine learning classifiers like - Naive Bayes, Decision Tree, Logistic Regression, and Random Forest- the recommended work predicts the chance of coronary heart sickness and classifies the affected person's threat level. Consensus modelling among the above-mentioned algorithms has been performed and overall accuracy of 91.25% is achieved.**

***Index Terms*—Cardiovascular Disease, Machine learning, IoT, Google Fit.**

I. Introduction

Over the last decade, cardiovascular diseases are the prime cause of death globally, taking approximately 17.9 million lives each year [1]. Heart diseases increase healthcare expenditures and lower wellness. According to the World Health Organization (WHO), India loses billions of dollars due to heart and cardiovascular diseases [2] [3] [4]. A range of factors, including personal and professional behaviors, as well as hereditary susceptibility, contribute to heart disease. Smoking, consumption of alcoholic beverages and caffeine, regular stress, and physical inactivity are all common risk factors for heart disease, as are health factors such as obesity, hypertension, high blood cholesterol, and preexisting cardiac illnesses [5]. The ability to diagnose heart disease rapidly, precisely, and accurately is essential for taking precautionary measures to avoid fatality. A general overview of the steps used for the detection of CVD is presented in Fig. 1. The correct analysis of the coronary heart sickness in patients is essential for reducing the associated risks of extreme coronary heart problems and enhancing the protection of the heart [6]. The invasive primary-based strategies mostly reason for vague prognosis and regularly put off with inside the prognosis effects because of human errors. Moreover, it is overpriced and computationally complicated and takes time to assessments [7].

Fig. 1: Overview of steps in detection of CVD.

II. Related Work

Prediction of heart diseases have been an active area of study. Varied Machine learning and Deep learning algorithms have been carried out for better chances of getting high accuracy. In [4], Rajdhan et al. tried out nearly all possible machine learning models that can be used to detect the presence of cardiac disease. The Evaluation metrics of Random Forest, Decision Tree, Logistic Regression, and Naive Bayes architectures that can be used for classification were compared in this study. In [7], Haq A et al. used seven well-known models, including logistic regression, Decision Trees, Artificial Neural Network, and 3 different feature selection techniques such as Relief, mRMR, and LASSO to select the best-related features. The K-fold cross-validation was also used. To evaluate the performance of classifiers, many assessment measures were used. The best accuracy was 89% when using the FS algorithm Reliwef's classifiers logistic regression with 10 - fold cross-validation. In [8], Ramalingam et al. examined the impact of feature extraction methods like Principal Component Analysis (PCA) on a variety of measures. Alternating decision trees have performed remarkably well when used with PCA, although decision trees haven't performed satisfactorily, probably due to overfitting. Ensemble models have done exceptionally well because they employ numerous algorithms to overcome the problem of overfitting. Models based on the Naive Bayes classifier were extremely fast in terms of computation and performance. SVM worked successfully in the great majority of situations. In [9], D. Shah et al. defined diversity of algorithms that can be employed to forecast heart disease effectively. K-nearest neighbor, Naive Bayes, and decision tree

978-1-6654-5342-4/22/$31.00 ©2022 IEEE

were used as data mining categorization approaches. The data was preprocessed before being incorporated into the model. The algorithms displayed were K-nearest neighbor (k = 7), Naive Bayes, and random forest.

III. Proposed Method

The research consists of 2 parts: Namely, (a) cardiovascular ailments prediction using Machine Learning/ Neural Networks Algorithms. (b) Predicting if a user is healthy enough based on fitness tracker data. The aim of part (a) is to classify people based on their habits to predict if they could have heart disease soon or not. Since this is a sensitive topic, to achieve the best results for the prediction, several different machine learning algorithms were employed and tested on the data. The heart disease dataset has been studied by various medical institutions to better understand the various factors that affect a person's heart's functioning. The models are compared with each other, and the results are shown. The series of events involves:

- Exploring the Framework
- Pre-processing of Data
- Machine Learning Algorithms

The work proposed is the implementation of algorithms, on a fitness tracker and smartwatch acquired data, to accurately predict the heart's condition.

A. Exploring the Framework

The system, shown in Fig. 1 consists of two parts. Namely, (a) compare the different machine learning algorithms and deploy the most suitable model. (b) used Django, HTML, CSS, JavaScript, and Firebase to build a web application. The system incorporates the use of Google Science-based metric "Heart Points". The system makes API calls along with the OAuth2 client to the Google Fit database using the Google developer console. The metric hence received is used to decide if a person is healthy or not.

1) Dataset Description: The dataset used in this study is combined and inspired from various datasets found under the Index of Heart Disease dataset from the UCI Machine Learning Repository. The major dataset is referred from the web repository of the University of California, Irvine. This dataset consists of results of 918 patients who presented for angiography at the Cleveland Clinic in Cleveland, Ohio. The Dataset has 11 impartial labels and a classification label (as shown in Table I), which is used to diagnose coronary heart sickness. The classification label has binary values, where 0 represents absence or relatively low threat level and 1 represents higher chances of disease. The data is partitioned into two splits of training and testing in the proportion of 80% and 20% respectively. The whole statistics of the dataset along with its description can be found in [10]. The metric along with some other information was stored in a vector. It is later subjected to Pre-processing for further refinement. Getting a bare minimum of heart points intermittently can assist someone to be healthy and free from any heart disease. Fig. 2 represents the proposed work structure, consisting of the ML/ Neural Networks structure.

TABLE I: Description of the features used.

S.No.	Feature	Label	Description	Domain
1	Age	AGE	Age in years	$29 < Age < 77$
2	Sex	SEX	Male=0 Female=1	1 0
3	Chest Pain Type	CP	1=Atypical Angina 2=Typical 3=Asymptomatic 4=NonAnginal Pain	787
4	Resting Blood Pressure	TRESTBPS	Blood Pressure in mm hg	$94 < rbp < 200$
5	Serum Cholesterol	SCH	In mg/dl	$120 < sch < 564$
6	Fasting Blood Sugar	FBS	$> 120 mg/dl$	1=True 0 = False
7	RestingECG Results	RESTECG	0 = Normal 1 = Having ST-T 2 = Hypertrophy	0 1 2
8	Max Heart Rate	THALACH	-	71-102
9	Exercise Induced Angina	EXANG	1=YES 0 = NO	1 0
10	Old peak	OLDPEAK	-	0.6.2
11	Slop of Peak exercise ST Segment	SLOPE	1=up sloping 2=Flat 3=down sloping	1 2 3
12	Target	TARGET	0=NO 1=YES	0 1

This section explains the resources, the datasets used, and the working of the proposed system.

2) Google FIT: The project involves a novel idea of using a Google-based metric for classifying whether a person is living a healthy life or not. The project uses the "Heart points" [11] or "Heart minutes" metric scientifically created by Google in association with American Heart Association's Centre for Health Technology & Innovation [12]. The extraction is done with the help of google fit API and few other technologies provided by Google for a safer connection between data flow. Since the application uses a person's private data, OAuth2 client has been established for privacy and security concerns.

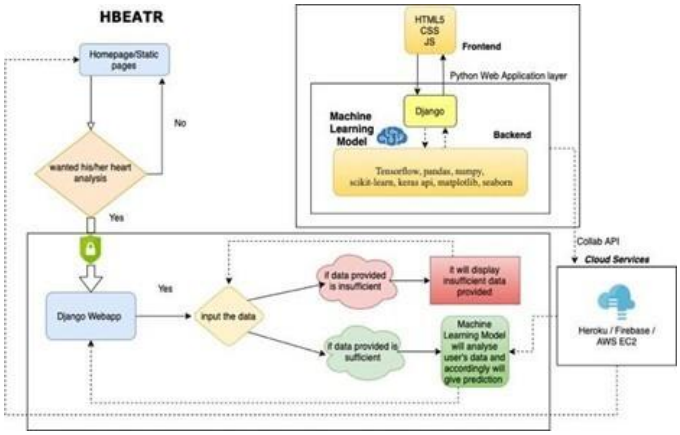

Fig 2. Overview of the proposed framework.

B. Pre-processing

The pre-processing of the dataset is necessary for visualization and efficient machine learning classification. A cleaned and processed data can lead to better training and validation of machine learning algorithms. Pre-processing techniques such as explanatory data analysis (EDA), handling missing values, and one-hot encoding of the categorical labels have been performed for better results. The data for the next part of the project, i.e., the data from Google FIT can be used directly, [13]. The data has been scaled and standardized as per the min- max scaling algorithm using the sklearn library in python. The data is scaled such that the data values are numeric and are in the range of 0 to 1.

C. Consensus of Classifiers

To perform the classification between heart disease patients and healthy people, Machine Learning algorithms/Neural Network models are used. Since the target variable has only 2 values 0 (Health) and 1(Abnormal) , binary classifiers are deployed [14]. The classifiers used are Random Forest Classifier, Support Vector Classifier (SVC), Artificial Neural Networks (ANN), Naive Bayes, and Logistic Regression. A consensus among these trained classifiers on same dataset is drawn with the help of Algorithm 1. In this algorithm Dataset being represented as T D. Initially the default label is set to healthy that turns to be abnormal in case the consensus result is obtained so. C represents the total number of classifiers considered for the purpose. The individual classifier models have been explained in the subsequent sections.

1) Logistic Regression: Logistic regression [15] is used to estimate the value of the result variable y when y is either 0 or 1,0 is the negative class and 1 is the positive class in a binary classification problem. Standardization of numerical values of the dataset is required as the model is linear and if all the numerical values are not within the same scale, then the bigger values could influence the performance of the model. The categorical data present is One-Hot Encoded. Categorical

Algorithm 1 Consensus of classifier(T D)

Require: T D $\neq \phi$
Ensure: label = T RUE
1: Read the dataset T D and set default label = Healthy
2: Initialize counter k = 0 > There are 5 classes considered
3: While k $\neq$ |C| l_k = Predict(C_k)
4: l = l $\cup$ l_k
5: Label = Majority(l)
6: End of While
7: Return Label

TABLE II: Performance metrics obtained for the models used.

Model	Accuracy	Sensitivity	Specificity	ROC
Logistic Regression	83.334%	82.30%	84.04%	83.99%
SVM	86.23%	80.53%	90.18%	85%
Naive Bayes	79.43%	94.33%	75.69%	80.73%
ANN	89.94%	86.98%	88.95%	84.74%
RF	85.86%	90.65%	94.49%	84.63%
Consensus	91.25%	92.05%	94.85%	89.68%

data tends to form different patterns in the dataset which could lead to unexpected model behavior.

2) Support Vector Machine: The proposed work uses a linear SVM model with soft margins. The dataset is processed similarly to the logistic regression approach. [16].

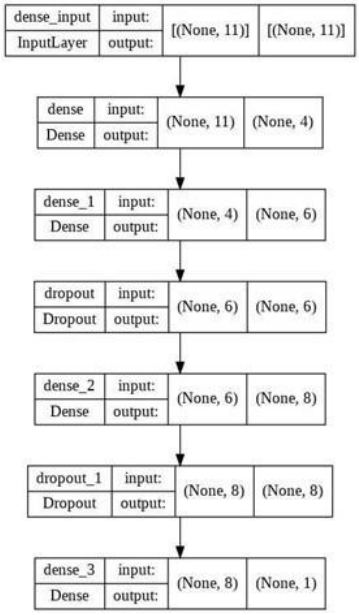

Fig. 3: Depicting the layers and number of the neural networks of the ANN deployed to test the classification.

3) Naive Bayes: The approach works best when attributes are highly independent in nature. [14] This assumption seldom works in the real world and hence, the approach was expected to be the worst performer.

4) ANN Stacking XGBoost: Artificial neural network being stacked with extreme gradient boost algorithm is considered

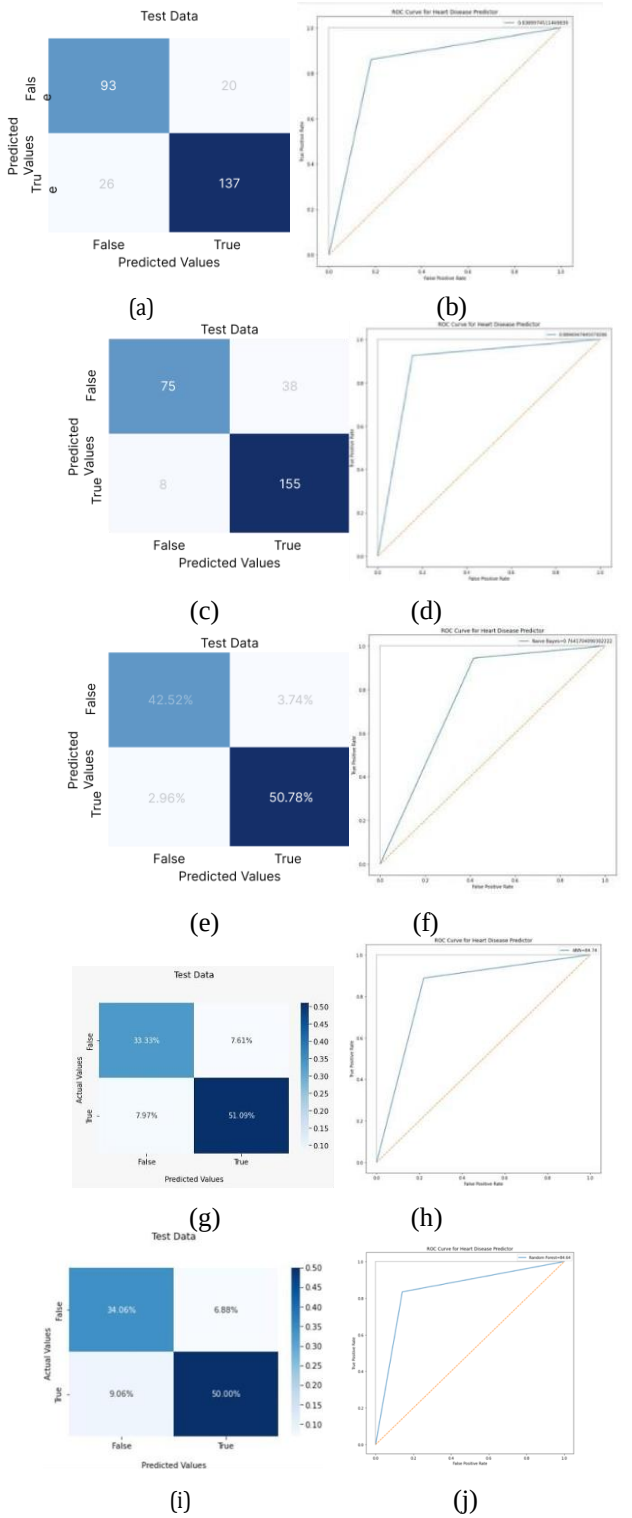

Fig. 4: Confusion matrix and ROC pairs for LR, SVM, NB, ANN, and RF in (a, b), (c, d), (e, f), (g, h), and (i, j) respectively.

for performing the said classification. Simply, the probabilistic outputs from the ANN are fed to the XGBoost module for an improved learning outcome. The ANN model consists of an input layer with 11 features. It has 2 hidden layers with 6 and 8 neurons, respectively. Both have a dropout of 0.3 with ReLU as the activation function. Finally, a single neuron output layer, with sigmoid as the activation function, gives the probability value of having a heart ailment. These probabilistic values are further used as input for the XGBoost algorithm which is stacked with this ANN model. For this experimentation, the data has been randomly split into training, validation, and testing sets in the ratio of 3:1:1. [17] The stacked model is trained for 100 epochs. A detailed structure of the ANN portion has been shown in Fig. 3.

5) Random Forest Classifier: Random Forests is a ”bagging” ensemble machine learning model that consists of various decision trees. It uses the concept of information gain and entropy to make decisions [18]. Both the Decision Trees and Random Forests are non-linear, lazy algorithms, Standard Scaling of the numerical labels is present in the dataset this time. The dataset is fed as it is to the model. The categorical data values are mapped into binary form for the model to understand them properly and not create any confusion for the model [19]. There are 2 hyper-parameters related to the Random Forests algorithm- the ”max-depth” parameter of the decision trees that are present in the Random Forests. The second parameter tuned is the ” n estimators” i.e., the number of trees in the forest. Both parameters are rigorously tested across different values using GridSearchCV techniques.

IV. RESULTS AND DISCUSSION

This section of the paper discusses the overall performance of the various classifiers in terms of the accuracy, sensitivity and specificity of these algorithms. The corresponding results of each classifier has been computed uniformly by utilizing GridSearchCV. K-fold (k = 5) cross-validation is used to measure the repeatability of the performance. Out of the above mentioned, it was observed that ANN stacking XG- Boost performed the best with the data. Further, a consensus approach enhances the accuracy performance at 91.25%. The probabilistic values used as input for the XGBoost algorithm showed a comprehensive outcome. Coming to the second part of the methodology, based on the research from the AHA’s Centre for Health Technology & Innovation, Heart Points data has been taken from Google Fit as it is a valid and scientific measure for checking if a person lives a healthy life. The value is generated because the Google FIT app will send Heart Points and Move Minutes at some point during walks, runs or motorcycle rides using smartphones or watch sensors, just like the accelerometer and GPS. Table 2 describes the different performance metrics the algorithm must represent for the evaluation. As shown in table 2, it is observed that the Consensus model works the best with the dataset. Sample confusion matrices and ROC’s for the models are presented in Fig. 4 in a pair-wise manner respectively for certain models as per their validations on the same sample set. A comparison

among competent schemes based on accuracy, sensitivity and specificity are also presented in Fig. 5.It is observed that the consensus outperforms the competent schemes

A. Pros and Cons

The sample distribution considered for validating the method is not even in nature. This makes some of the models perform averagely efficient as in case of regression. Further, the loosely related features effects the efficiency of the Naive Bayes'. Non-linearity and laziness cannot be exploited by the uniform intervals in case of random forest. However, despite of certain cons, the consensus of these models results in a compensatory basis performance that is quite satisfactory.

Fig. 5: Comparing the performance scores among competent schemes.

CONCLUSION AND FUTURE SCOPE

Prediction of cardiovascular ailments through computational models is well validated through a series of experiments on suitable datasets. Along with this the importance of the API feed parameter "heart points" is realized. A modified Consensus model is observed to be outperforming competent state- of-the-art models in terms of performance measures. The experimentations so carried out show an overall rate of accuracy of 91.25 along with sensitivity at 92.05 in favor of the Consensus model. It can also be concluded that heart points parameter can be a viable data point which can be served as a measure for classification of people's health habits However analysis is to be carried out as future work on a way to cope with excessive dimensional facts and overfitting. Subsequent studies can be executed at the quality set of rules ensemble to rent for a specific kind of facts.

REFERENCES

[1] Mensah G, Roth G, Fuster V, et al. The Global Burden of Cardiovascular Diseases and Risk Factors. J Am Coll Cardiol. 2019 Nov, 74 (20) 2529–2532. https://doi.org/10.1016/j.jacc.2019.10.009

[2] Gerc V, Masic I, Salihefendic N, Zildzic M. Cardiovascular Diseases (CVDs) in COVID-19 Pandemic Era. Mater Sociomed. 2020 Jun;32(2):158-164. doi: 10.5455/msm.2020.32.158-164. PMID: 32843866; PMCID: PMC7428924.

[3] Nayak, G., Padhy, N. Mishra, T.K. 2D-DOST for seizure identification from brain MRI during pregnancy using KRVFL. Health Technol. 12, 757–764 (2022). https://doi.org/10.1007/s12553-022-00669-4

[4] ApurbRajdhan, Avi Agarwal, Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020)

[5] Gerc V, Masic I, Salihefendic N, Zildzic M. Cardiovascular Diseases (CVDs) in COVID-19 Pandemic Era. Mater Sociomed. 2020 Jun;32(2):158-164. doi: 10.5455/msm.2020.32.158-164. PMID: 32843866; PMCID: PMC7428924.

[6] B. Alić, L. Gurbeta and A. Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases," 2017 6th Mediterranean Conference on Embedded Computing (MECO), 2017, pp. 1-4, doi: 10.1109/MECO.2017.7977152.

[7] Haq A, Li JP, Memon MH, Nazir S, Sun RA (2018) hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Inf Syst 2018:1-21.

[8] V. RAMALINGAM, V; DANDAPATH, Ayantan; KARTHIK RAJA, M. Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, [S.l.], v. 7, n. 2.8, p. 684-687, mar. 2018 .

[9] Shah D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020- 00365-y.

[10] Rani, P., Kumar, R., Jain, A. (2021). Multistage Model for Accurate Prediction of Missing Values Using Imputation Methods in Heart Disease Dataset. In: Raj, J.S., Iliyasu, A.M., Bestak, R., Baig, Z.A. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore. https://doi.org/10.1007/978-981-15-9651-3_53

[11] Mehdi Nobakht, Yulei Sui, Aruna Seneviratne, Wen Hu, PGFit: Static permission analysis of health and fitness apps in IoT programming frameworks,Journal of Network and Computer Applications,Volume 152,2020,102509,ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2019.102509.

[12] Muntner, Paul, Daichi Shimbo, Robert M. Carey, Jeanne B. Charleston, Trudy Gaillard, Sanjay Misra, Martin G. Myers et al. "Measurement of blood pressure in humans: a scientific statement from the American Heart Association." Hypertension 73, no. 5 (2019): e35-e66.

[13] Behnami, D. (2022). Machine learning for diagnosing functional heart disease in echocardiography (Doctoral dissertation, University of British Columbia).

[14] X. Wu and V. Kumar, Top 10 Algorithms in Data Mining,Springer, Berlin, Germany, 2007.

[15] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machinelearning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4): e0174944.

[16] Vishwanathan, S. V. M., &Murty, M. N. (2002, May). SSVM: a simple SVM algorithm. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290) (Vol. 3, pp. 2393- 2398). IEEE.

[17] Taghizadeh-Mehrjardi, Ruhollah, Karsten Schmidt, Alireza Amirian-Chakan, Tobias Rentschler, Mojtaba Zeraatpisheh, Fereydoon Sarmadian, Roozbeh Valavi, Naser Davatgar, Thorsten Behrens, and Thomas Scholten. 2020. "Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space" Remote Sensing 12, no. 7: 1095. https://doi.org/10.3390/rs12071095

[18] Kryvenchuk, Yurii, Alina Yamniuk, Iryna Protsyk, Lesia Sai, Andriana Mazur, and Olena Sydorchuk. "Random Forest as a Method of Predicting the Presence of Cardiovasculars Diseases." (2022).

[19] Mohan Kumar, K.N., Sampath, S., Imran, M., Pradeep, N. (2021). Clustering Diagnostic Codes: Exploratory Machine Learning Approach for Preventive Care of Chronic Diseases. In: Satapathy, S., Zhang, YD., Bhateja, V., Majhi, R. (eds) Intelligent Data Engineering and Analytics. Advances in Intelligent Systems and Computing, vol 1177. Springer, Singapore.