International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hybrid Machine Learning Model for Chronic Disease Prediction

Rahama Salman *[1], Dr. Subodhini Gupta[2]

Submitted: 10/11/2022 **Revised:** 15/01/2023 **Accepted:** 08/02/2023

Abstract: The previously suggested chronic disease prediction techniques are incapable of acquiring efficiency in feature extraction, outlier removal and classification. This research work is conducted to tackle the limitations of these methods. After eliminating the existing drawbacks, the accuracy to predict the chronic disease is augmented consequently. Therefore, the fundamental emphasize is on predicting the disease on the basis of economic and social data, and analyzing the trends of chronic diseases depending upon the epidemiological data. This work suggests a hybrid framework in which Random Forest (RF) is integrated with Logistic Regression (LR). The initial algorithm is implemented for extracting the features, and the latter one is exploited for classifying the diseases. Logistic Regression algorithm makes the deployment of extracted features as input to classify the data. Python is executed to simulate the suggested framework. Various metrics, namely accuracy, precision and recall are utilized to analyze the results.

Keywords: *Chronical Disease, Machine Learning, Hybrid Model, Logistic Regression, Random Forest*

1. Introduction

Chronic disease-related health issues are becoming more severe as modern society ages. Even without risk factors, the absolute risk levels for cardiovascular disease increase with age, although the risk variables vary depending on the age group [1]. Because physical risk factors including lipid metabolism and vascular status in young individuals do not significantly affect cardiovascular disease (CVD), the incidence of cardiovascular disease amongst youngsters is smaller than that of other age categories [2]. Physiological risk factors had a more gradual effect on middle-aged people's cardiovascular disease risk factor features than they did on young people, but it took some time before the risk factor dramatically worsened. As a result, in comparison to the elderly, who are the actual risk category for cardiovascular disease, the middle-aged are simply considered to be a potential risk group [3]. This suggests that the middle-aged, who are the present potential risk group, are quite likely to get cardiovascular disease if they continue to lead an unhealthy lifestyle they have had since they were young. As a result, it is crucial to identify and treat those who are at high risk for cardiovascular disease in advance. It would also be highly helpful to have a technology that could identify when the disease will start [4].

Data-oriented cardiac diagnostics fall within the definition of Machine Learning (ML), which is the application of computer algorithms with the capacity to learn to carry out specific actions from example data without needing publicly programmed instructions [5]. In order to produce the most accurate predictions utilizing the original data, this branch of Artificial Intelligence (AI) uses sophisticated statistical techniques to extract predictive or preferred patterns from training data [6]. An effective ML model is reliant on the availability of relevant and accurate data. As a result, gathering the right data is essential to ensuring that the ML model can execute both internal and external authentication with great efficiency [7]. Machine learning models that recognize disease according to the risk level can identify subjects with a high likelihood of having CVD [8]. After that, interventions like healthy lifestyle plans tailored to this target population and the dissemination of information can reduce the risk of growing CVD and support ground-breaking treatments for cardiovascular illnesses. A set of regulations on the assessment and management of CVD recommends using enforced detection algorithms to identify populations at higher risk, and make the clinical decision.

1.1. Machine Learning Techniques for CVD Prediction

Machine learning (ML), a subtype of AI, is becoming more and more accepted in the field of cardiovascular medicine. Software systems can become increasingly accurate at predicting outcomes thanks to machine learning (ML), without having to be specifically programmed to do so. Verifiable data is used as a contribution by machine learning algorithms to forecast new result values. A model that receives input data (such as images or text) and predicts outcomes (such as ideal, problematic, or unbiased) by combining numerical improvement and quantifiable research is the foundation of the hypothetical machine learning architecture. Data-based CVD diagnostics are an

1 Research Scholar, Department of Computer Science, guddu.rahama@gmail.com
ORCID ID: 0000-0003-4105-5019
2 Associate Professor, Department of Computer Application, School of Information Technology,subodhinig@gmail.com
1,2 SAM Global University, BHOPAL, MP, INDIA
** Corresponding Author Email: guddu.rahama@gmail.com*

example of an application of machine learning (ML) that uses computerized algorithms to discover how to carry out selected assignments using model data without the need for clear automated instructions [9]. This branch of AI employs cutting-edge quantifiable techniques to identify fair or predictive instances from preparing data to forecast outcomes with the highest degree of accuracy on the obtained data. The following describes the typical machine learning techniques for the detection of cardiovascular events.

1.1.1 . Logistic Regression (LR)

It is able to effectively predict the probability of a given result depending on input variables continuously, unlike a binary classification system. The final probability is multiplied by 1. As a result, one gains definition for every possible result and everyone's prospects. The forecasting of conclusive likelihood is negatively impacted by the slightest informational difference, which is another characteristic of this approach. Additionally, it is vital to keep the information vector aspect low because doing so increases the cost of model preparation and the risk of the model being overfit to the preparation dataset. Principal component (PCA) [10] or LDA are used to moderate the amount of predictors to those that provide more convenience in the case of a large amount of information factors. The algorithm LR is suitable when combining data from many sources in a binary classification task and requires less complexity.

1.1.2 Support Vector Machine (SVM)

It is a supervised ML approach which determines the ideal linear or non-linear limit by classifying the input into two or more groups. The most crucial function, that of bit determination, has been completed because it is this function that isolates the data. It is easy to access the broad bits as a linear function. The training of several frameworks is done, and the settings are preserved to select the other SVM measures for the algorithm with the lowest error. This algorithm can't separate aspects well enough. Thus, in order to simplify the training and achieve a more accurate LR speculation, a strategy of lowering the focus on the information variables is used. The main limitation of this technique is that it uses more memory while processing large amounts of data. The non-linearity and sparsity in the information data are successfully perceived by this algorithm. SVM is a complex method, and several studies enable its application to achieve the highest viability [11].

1.1.3 Random Forest

With numerous noisy though roughly unbiased algorithms, bagging is a strategy which focuses on minimizing the variance of an estimated predictor. An assortment of uncorrelated decision trees makes up a random forest. A training set is created for each decision tree by randomly selecting and replacing a section of the source data. A random subset taken from the collection of attributes is employed in this algorithm at every node split for generating each decision tree. The algorithm essentially employs bagging for trees as well as attributes. There is no pruning; rather, each tree is allowed to achieve its maximum growth [12]. The average of several trees, till the trees are uncorrelated, is more prone towards noise in its training set, whereas a single tree's forecasts are. By building the trees utilizing various training sets, bagging accomplishes de-correlating. Random Forests use the results of the developed trees in the hybrid to produce a prediction for a fresh sample. For big data sets, random forests operate very effectively, are capable of handling data sets with imbalanced classes, and do not run the risk of overfitting (unlike, for example, the boosting method AdaBoost). Through cross validation, the ensemble's number of trees is chosen.

1.1.4 . Artificial Neural Network (ANN)

It is designed according to the composition and relations of Neural Networks (NNs) with biological inspiration. In order to insert input data into the hierarchy's various layers, internal nodes are used. Each input line contains an associated weight whose evaluation is done, and it is modified when the data is trained. Artificial neural networks (ANNs) are used until exceptional results-producing weights are found. Each node's nonlinear function is used to calculate the value and take into account the input from incoming connections [13]. Because weight optimization uses highly non-linear combinations of these algorithms' properties, algorithms become more flexible to difficult limits. This algorithm also considers the relationship between the layers in order to construct various networks depending on the application. In accordance with every issue and the volume of data used to train the model, the current hypothesis places some restrictions on several motes in every layer, and such algorithms support a propensity to adjust to the training phase. The algorithm's weights are different from those in the training samples, which explains why. ANN worked well even when there was a lot of data there.

2. Literature Review

A. Rahim, et.al (2021) introduced a MaLCaDD algorithm in order to predict cardiovascular disease (CVD) at higher precision [14]. In general, Particularly, the mean replacement (MR) method was employed to tackle the missing values and Synthetic Minority Over Sampling Technique (SMOTE) algorithm assisted in addressing the issue of data imbalance. Afterward, Feature Importance (F1) method employed to select the attributes. At last, an ensemble of Logistic Regression (LR) and K-Nearest Neighbor (KNN) was put forward to predict CVD at superior accuracy. The introduced model yielded an accuracy up to 99.1% on Framingham, 98.0% on HD dataset and 95.5% on Cleveland dataset. Furthermore, the reliability and applicability of the introduced model was proved in real environment. A study by H. D. Park et al. (2018) introduced a technique that predicts the risk of cardiovascular disease (CVD) based on the disease's attributes [15]. A new framework called FA-Attn-LSTM was suggested for this purpose. FA-Attn was exploited for inserting weights to input data at higher correlation amid the input data and predictive object. The data gathered from the Asan medical center applied for computing the suggested framework in the experimentation. The results confirmed the robustness of the suggested framework. Moreover, this framework attained root means square error (RMSE) around 3.65 and mean absolute error (MAE) of 2.49. The suggested technique was adaptable on the healthcare data as well as in other fields with asynchronous data. Y. An et al. (2021) introduced DeepRisk, a fully end-to-end (E2E) algorithm that relies on an attention system and a Deep Neural Network (DNN) algorithm. The algorithm is designed to automatically learn high-quality features from Electronic Health Records (EHRs) and combine heterogeneous data with time-ordered clinical data in an automated manner [16]. A real time clinical dataset was executed in the experiments for evaluating the developed approach. According to the experiments, the developed approach helped in enhancing the accuracy of predicting the high-risk CVD as compared to the existing techniques. In their study, W. Zeng et al. (2021) utilized a stepwise regression analysis (SRA) approach with the Akaike Information Criterion (AIC) to identify the relevant variables associated with cardiovascular disease (CVD) [17]. An artificial neural network (ANN) based survival framework was implemented for predicting CVD. A comparative analysis was conducted on ANN against SVM and NB. The outcomes exhibited the supremacy of the projected approach over other methods. The strongly correlated variables were utilized to enhance the efficacy of every algorithm. The projected approach offered an accuracy of 81%. Recall of 83%, F1-score of 84% and area under curve (AUC) of 84%. In their research, D. M. R. et al. (2022) explored a hybrid framework combining Binary Particle Swarm (BPS) and Constrained Optimization (CO). The goal of this framework was to optimize the number of created clusters and maximize efficiency, resulting in a reduction in time complexity [18]. ROUGUE technique applied for quantifying this framework. Based on the experiments, the investigated framework attained superior precision, recall and F-Scores in comparison with other methods. Furthermore, this technique was useful for alleviating the entropy and time complexity while predicting cardiovascular disease (CVD). K. Junwei, et.al (2019) emphasized on predicting cardiovascular disease (CVD) based on enhanced Long Short-Term Memory (LSTM) algorithm [19]. The conventional method was considered to generate this algorithm for predicting CVD. The time parameter vector was acquired after its ITI was smoothened, and its deployment was done as input of the forgetting gate of Long Short-Term Memory for tackling the predictive obstacle occurred due to ITI. The findings demonstrated the supremacy of the presented algorithm over the existing techniques. The irregular time among diverse clinical phases of the patient was smoothed to enhance the presented algorithm. In their study, V. S. Dehnavi et al. (2020) created an intelligent system for evaluating the risk of cardiovascular disease (CVD) [20]. Initially, they employed a neuro-fuzzy network (NFN) as a predictive model to assess the risk of CVD. The data of patient, including blood sugar (BP), heart rate (HR), smoking, age, was exploited as input and the output provided the risk factors of CVD for patients over upcoming decade. Genetic algorithm (GA) was implemented to mitigate the attributes used to determine the condition of patient. The least-square (LS) algorithm helped in verifying the metrics of network. The nonlinear metrics of fuzzy sets were optimized through improved grasshopper optimization algorithm (IGOA). In the end, the designed system was quantified on the Framingham dataset. The simulation algorithms validated that the acceptability of the designed system. P. Theerthagiri, et.al (2022) discussed that machine learning (ML) models often assisted in predicting the risk indicators of cardiac disorders [21]. Thus, RFE-GB model was recommended to predict the cardiac disorders. The health records of individuals suffered from cardiovascular disease (CVD) employed to compute the results. various ML techniques were implemented to construct this algorithm and perform its comparison. The accuracy of the recommended algorithm was calculated 88.8% and area under the curve (AUC) was 0.84 in contrast to the existing techniques. Additionally, the recommended algorithm was effective tool to predict and cure CVD. In their research, R. Li et al. (2021) devised an approach for predicting cardiovascular disease (CVD) by integrating the logistic regression (LR) technique and similarity evaluation (SE) utilizing a distance formula [22]. Kaggle platform was considered to derive the data. A first-stage model was generated using LR. The formulated method attained an accuracy of 88.5% and recall of 89.2%. SE technique relied on Euclidean distance (ED) was utilized for optimizing the data whose processing was done in the initial phase. In the end, a complete classifier was constructed. This classifier provided an accuracy of 98% and recall of 98% for predicting CVD. In their study, A. Elbadry et al. (2021) put forward the MOMESRA technique, which involved the inclusion of various electrocardiogram (ECG) segments from a single patient and utilizes a majority-vote mechanism to predict the risk of cardiovascular disease (CVD) [23]. This approach focused on sub-dividing the electrocardiogram segments into 2-minute windows which were utilized to extract the attributes such as TD, FD and nonlinear HRV. (Principal Component Analysis (PCA) algorithm deployed to mitigate the attributes. Support Vector Machine (SVM) technique employed alleviated attributes as input. The suggested technique yielded 78% accuracy, 90% sensitivity, 67% specificity and 79% precision. The outcomes demonstrated the effectiveness of the suggested technique to predict CVD events. N. S. Rajjliwal, et.al (2021) devised a framework on the basis of SD-CNN-C which predicted cardiovascular disease (CVD) robustly and resiliently against the class imbalance [24]. First of all, LASSO was employed and the majority voting stage was executed to extract the crucial and homogenized attributes. after that, a devised framework and a new model development protocol was adopted. A particular training routine per epoch was exploited for enhancing the efficacy and robustness for dealing with class imbalance. The experimental outcomes on NHANES data indicated that the devised framework offered an accuracy up to 0.818 for normal cases and 0.85 for infected cases.

3. Research Methodology

Chronical disease results in creating a menace to human life due to its contagious nature, and causing a greater mortality rate. The complex task is of diagnosing and curing such diseases at their premature phases. The inadequate access to diagnostic services, professional medical personnel, and other resources results in making the process of prognosis the disorders more intricate. Any kind of disease leads to lay impact on other bodily organs and create disturbance. For predicting the heart condition, researchers often prefer to implement computer-aided data extraction from an enormous amounts of records. Data Mining (DM) algorithms and procedures attain popularity among various organizations. The DM algorithm renders higher accuracy for predicting chronical disease in the healthcare sector. Such kind of disorders may occur because of a variety of risk factors. Chronical diseases are predicted in diverse phases which are elucidated as:

3.1. Data Gathering

This phase emphasizes on assembling the data taking various clinical associations into consideration. The collected data is deployed for testing the presented approach.

3.2. Data Pre-Processing

The fundamental objective of this phase is to analyze the data and attain the completeness. For this, Machine Learning (ML) methods are implemented. Furthermore, the data is pre-processed with the intend of eradicating the redundant features from the dataset. Accordingly, the further task is to enhance the training system in order to clean and de-noise the data whose dissemination is done in the succeeding phase.

3.3. Feature Selection

A subgroup with distinctive qualities is expedient to diagnose the cardiac ailments. These attributes are proficient in coping with the current class of attributes. Random Forest (RF) algorithm is implemented for selecting the features. This algorithm assigns a value of 100 to the estimator value. In this, the construction of a tree structure is performed in which all of the relevant features are encompassed. The modification of this algorithm is assisted in selecting the pertinent features to diagnose the coronary diseases.

3.4. Classification

This phase is executed for mapping the chosen attributes onto the training system such that the features, deployed for input, are classified. Consequently, this stage acquires the effectiveness of predicting the coronary issues. Every discrete class is responsible for demonstrating a certain category of coronary illness. The collected characteristics are exploited as input in the Logistic Regression (LR) algorithm for classifying the chronical diseases. For specifying the probability of an individual having a crucial heart condition, this research effort leads to split the data into two categories: heart disease and healthy.

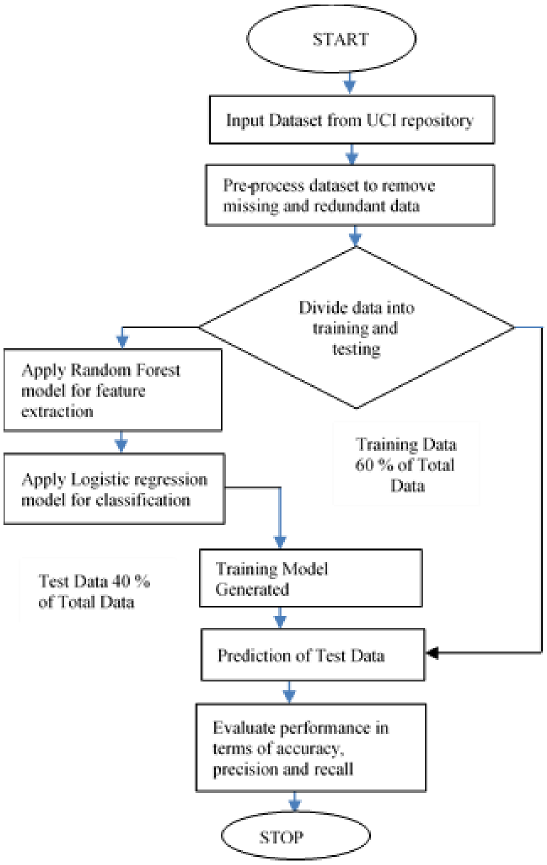

START
Input Dataset from UCI repository
Pre-process dataset to remove missing and redundant data
Divide data into training and testing
Apply Random Forest model for feature extraction
Apply Logistic regression model for classification
Training Data 60 % of Total Data
Test Data 40 % of Total Data
Training Model Generated
Prediction of Test Data
Evaluate performance in terms of accuracy, precision and recall
STOP

Fig 1. Proposed Model

3.4.1. Proposed Hybrid Machine Learning Model

Hybrid machine learning (ML) is an approach in which diverse ML methods are integrated for attaining their potential at individual level and tackling their drawbacks. A number of methods namely decision trees (DTs), and neural networks (NNs), etc. are put together for generating a hybrid approach in order to attain higher efficiency in comparison with a single technique. This approach is capable of dealing with complicated issues which are not resolved with the help of individual algorithm. Different domains, such as finance, healthcare, and marketing, adopt hybrid ML to accomplish diverse objectives.

In the given figure 1, the process of detecting cardiovascular disease (CVD) is defined. This work suggests a hybrid technique of Random Forest (RF) and Logistic Regression (LR) for detecting CVD. First of all, UCI repository is considered to collect the CVD data. Thereafter, the data is pre-processed for eliminating the misplaced and inadequate data. The next step is to divide this data into 2 sections: training and testing. At first, the RF algorithm is adopted to extract the features. This algorithm also assists in selecting the attributes. These algorithms are planned on the basis of tree based techniques and the ranks are assigned to them according to their efficiency for maximizing the node purity. It leads to illustrate an overall alleviation in impurity. At the initial stage of the trees, the nodes are available that have decrease in impurity. However, at the end, trees are consisted of nodes having the smallest reduction in impurity. Hence, a subset of significant attributes is created when the trees are cut below a given node. Thereafter, the LR algorithm is implemented for classifying the data. Logistic Regression is a widely-used supervised machine learning binary classification algorithm, known for its simplicity and effectiveness. It is particularly useful for working with categorical dependent variables, with the output being a discrete or binary categorical variable (i.e., 0 or 1). The algorithm employs the sigmoid function as its cost function, which maps predicted real values to a probabilistic value between 0 and 1. This approach enables the algorithm to provide a reliable prediction of the probability of an event occurring, based on input variables. The logistic Sigmoid function can be defined as:

$$P(x) = 1/(1 + e \wedge (-x))$$

In the logistic regression algorithm, the probability estimation function P(x) is used to provide a predicted probability value between 0 and 1. The input to this probability function is represented by the algorithm's prediction value, x. The function makes use of the mathematical constant e, which is Euler's number and has an approximate value of 2.71828, as shown in above equation. To predict the occurrence of cardiac disease, a logistic regression machine learning model is commonly employed. The model is first trained using five different splitting conditions, and then tested with a separate dataset to determine its predictive accuracy and behaviour. The algorithm outputs a category of either 1 or 0, indicating the presence or absence of cardiac disease, respectively. The next stage focuses on generating the trained model. moreover, the testing data is predicted. Diverse metrics are considered to compute the performance of the suggested technique. Hence, the CVD is predicted effectively. The second way after splitting the data is to predict the testing data its features are already extracted and evaluate the performance.

4. Result and Discussion

This work emphasizes on predicting the chronical disease. Three datasets are considered for the results analysis. The datasets are heart disease, diabetes disease and kidney disease.

The heart disease dataset is collected from https://archive.ics.uci.edu/ml/datasets/heart+disease which has seventy-six attributes. However, a subset of 14 of them is utilized in all published experiments. The "goal" field is employed for illustrating the occurrence of coronary disease in the patient. It is depicted numerically within 0 to 4.

The second dataset is of diabetic dataset which is generated via the source https://archive.ics.uci.edu/ml/datasets/diabetes. two sources called an automatic ER device and paper records are considered to retrieve the records of diabetic individuals. A clock is inserted in the automatic device for computing timestamp events, and the paper records helps in offering the "logical time" slots such as breakfast, lunch, dinner, and bedtime. In latter paper records, a definite time is allocated to each slot such as 08:00, 12:00, 18:00, and 22:00 respectively. Therefore, the initial records are useful to attain real-time time stamps, and the latter offer fictional uniform recording times.

The third dataset is of kidney dataset which is collected from https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. This data is collected within the duration of two months in India and twenty-five attributes, such as red blood cell count, white blood cell count, etc., are comprised in this. The data is classified in two classes: CKD or normal. The dataset is consisted of 400 rows.

The proposed model is implemented and results of the technique is compared concerning accuracy, precision and recall. The detail description is the parameters is provided below: -

4.1.1. Accuracy

This parameter is computed by dividing the positively classified cases with total amount of cases and multiplying them with 100.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

4.1.2 Precision

In the process of recognizing pattern, information is retrieved and binary classification is done and precision is computed by dividing the positively classified cases with all the cases.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

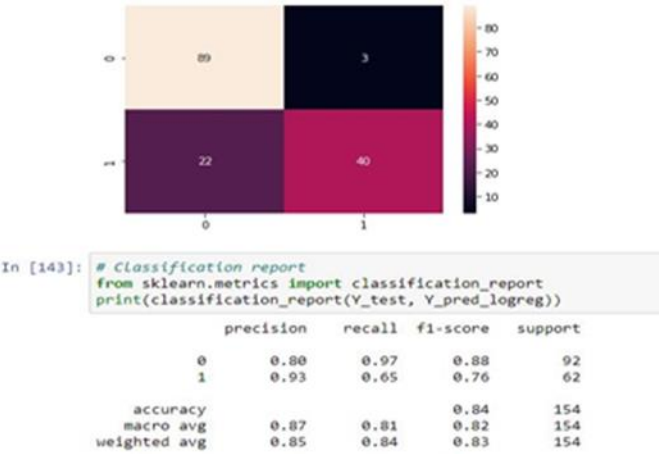
4.1.3. Recall

It is evaluation after dividing the positive instances which are retrieved with the total amount of all the cases.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

4.2 Diabetes Prediction Results

The three dataset of chronical disease is used for the prediction. The diabetes is one of the cardiovascular diseases. The proposed model is implemented for the diabetes prediction and compared with other classification models like LR, KNN, SVM, NB, DT and RF in terms of accuracy, precision and recall.



```
In [143]: # Classification report
          from sklearn.metrics import classification_report
          print(classification_report(Y_test, Y_pred_logreg))

              precision    recall  f1-score   support

           0       0.80      0.97      0.88        92
           1       0.93      0.65      0.76        62

    accuracy                           0.84       154
   macro avg       0.87      0.81      0.82       154
weighted avg       0.85      0.84      0.83       154
```

Fig. 2. Proposed Model Execution

Figure 2 depicts that the suggested framework execution is shown on the diabetic dataset.

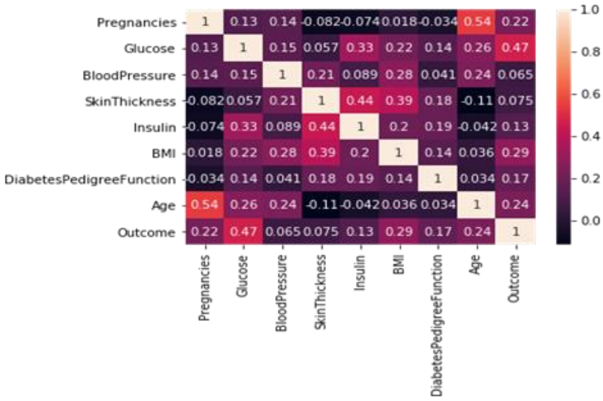


Fig. 3. Attributes of Dataset

As shown in figure 3, various attributes of dataset are plotted along with the relation of attribute with other attribute.

TABLE I. Performance Analysis on Diabetes Dataset

Model	Accuracy	Precisio n	Recall
Logistic Regression	71.42	73	72.2
KNN	78.37	79	78
SVM	73.37	75	74
Naïve Bayes	71.42	72	71
Decision Tree	68.18	69	68
Random Forest	75.17	76	75
Proposed Model	83.76	85	83

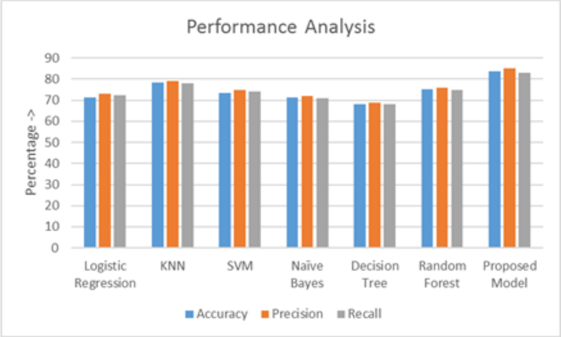
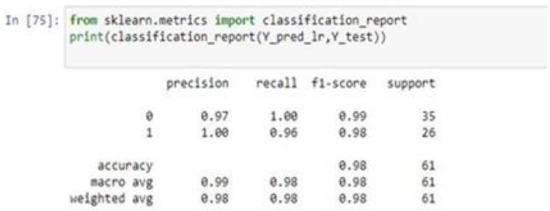


Fig. 4. Performance Analysis on Diabetes Dataset

As shown in figure 4, the comparison of the performance of proposed framework is done with the existing methods for the diabetic prediction. The analysis exhibited that the suggested framework achieves accuracy, precision and recall up to 84 percent approx. which is 10 percent higher than other classification models.

4.3 Heart Disease Prediction

The second major chronical disease is coronary disease. The testing of the suggested framework is done on the heart disease prediction. A comparative analysis is conducted on the proposed model against other models concerning accuracy, precision and recall.



```
In [75]: from sklearn.metrics import classification_report
         print(classification_report(Y_pred_lr,Y_test))

              precision    recall  f1-score   support

           0       0.97      1.00      0.99        35
           1       1.00      0.96      0.98        26

    accuracy                           0.98        61
   macro avg       0.99      0.98      0.98        61
weighted avg       0.98      0.98      0.98        61
```

Fig.5. Proposed Model on Heart Disease Dataset

Figure 5 depicts that the implementation of the suggested model on heart disease dataset.

TABLE II. Performance Analysis of Heart Disease Dataset

Model	Accuracy	Precision	Recall
Logistic Regression	85.25	86	85
KNN	67.21	67	67
SVM	81.97	83	81
Naïve Bayes	85.25	85	84
Decision Tree	81.97	82	81
Random Forest	95.08	96	95
Proposed Model	98.37	98	98

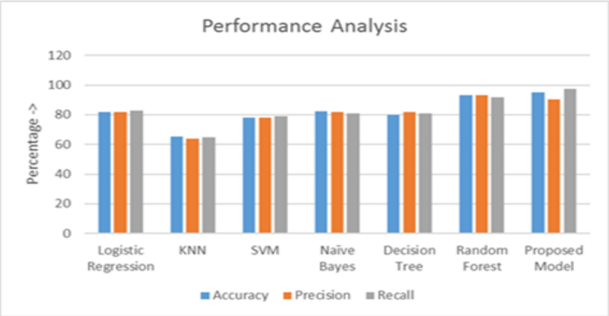


Fig.6. Performance Analysis on Heart Disease Dataset

As shown in figure 6, the suggested framework is tested on heart disease dataset. The proposed model achieves maximum of accuracy of 98 percent which is approx. 3 percent higher than other models.

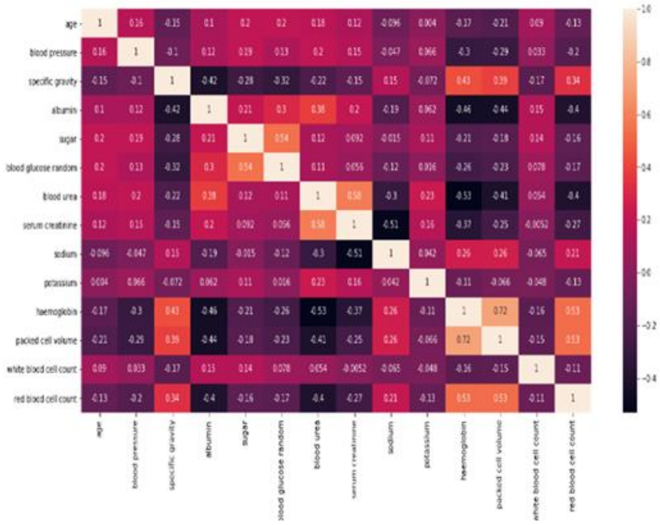


Fig.7. Attributes of Dataset

As shown in figure 7, various attributes of kidney disease are shown in the form of confusion matrix. The matrix displays relationship of attributes with each other.

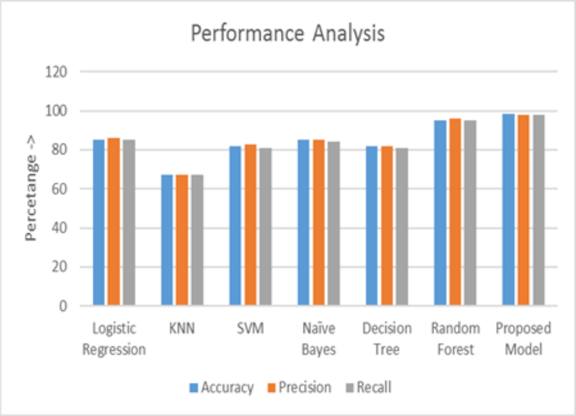


Fig.8. Performance Analysis on Kidney Disease

TABLE III. Performance analysis for Kidney Disease Prediction

Model	Accuracy	Precision	Recall
Logistic Regression	82	82	83
KNN	65.09	64	65
SVM	78.18	78	79
Naïve Bayes	82.16	82	81
Decision Tree	80	82	81

Random Forest	93.02	93	92
Proposed Model	95	90.25	97.34

As shown in figure 8, the suggested framework is tested for predicting the kidney disease. It is analysed that suggested framework achieves accuracy of 95 percent which is 2 percent higher than existing model.

5. Conclusion

In recent times, the foremost concern regarding the chronic disease is that there is not any symptom appearance in the patient who agonizes from chronic ailments. Such an infected patient leads to communicate this disease and infested other individuals all around the globe. Innumerable reasons are present for the occurrence and progression of chronic diseases. The first reason is the deficiency of information transparency in the premature stage of this epidemic outbreak. Distinct phases are executed for the prognosis of chronic disease such as to pre-process the data, extract the features and classify the data. This work suggests a hybrid framework in order to predict the chronical disease. It is an amalgamation of Random Forest (RF) and Logistic Regression (LR) algorithms. RF algorithm is implemented for extracting the features and LR algorithm assists in categorizing the diseases. The testing of the suggested framework is performed on three dissimilar disease datasets consisted of diabetes, kidney and heart disorder. Different metrics such as accuracy, precision and recall are considered for quantifying the suggested framework. According to results, the suggested framework performs more effectively on all datasets in comparison with the traditional models.

Acknowledgements

This research is an original work depicting disease prediction using ML based algorithms. I am extremely grateful to my supervisor Dr.Subodhini Gupta who provided insight and expertise that greatly assisted the research. I thank Dr.Neelesh Jain, Director at SGU for assistance with new techniques and trends in the technology and Dr.Akhilesh Kumar Singh, Dean R&D ,SGU for comments that greatly improved the manuscript.

Conflict Of Interest

"The authors declare no conflict of interest".

Author Contributions

Author Rahama Salman made a majority of the contribution to the Study conception and design, data collection, proposed research methodology, analysis and interpretation of results, and manuscript preparation and Author Subodhini Gupta made a substantial contribution to the concept, design of the article, interpretation of data for the article and approved the version to be published. Both authors have approved the final version.

References

[1] M. U. Khan, S. Zuriat-e-Zehra Ali, A. Ishtiaq, K. Habib, T. Gul and A. Samer, "Classification of Multi-Class Cardiovascular Disorders using Ensemble Classifier and Impulsive Domain Analysis," 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), 2021, pp. 1-8

[2] N. Rajesh, A. C. Ramachandra and A. Prathibha, "Detection and Identification of Irregularities in Human Heart Rate," 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-5

[3] S. M. Rayavarapu, D. Bikshapathi, S. L. Sabat and J. S. A. E. Fouda, "FPGA implementation of Ordinal Pattern Analysis algorithm for Early Detection of Cardiovascular diseases," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-4

[4] R. Banerjee, S. Bhattacharya, S. Bandyopadhyay, A. Pal and K. M. Mandana, "Non-Invasive Detection of Coronary Artery Disease Based on Clinical Information and Cardiovascular Signals: A Two-Stage Classification Approach," 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 2018, pp. 205-210

[5] B. D. Sekar, M. Chui Dong, J. Shi and X. Y. Hu, "Fused Hierarchical Neural Networks for Cardiovascular Disease Diagnosis," in IEEE Sensors Journal, vol. 12, no. 3, pp. 644-650, March 2012

[6] O. Terrada, A. Raihani, O. Bouattane and B. Cherradi, "Fuzzy cardiovascular diagnosis system using clinical data," 2018 4th International Conference on Optimization and Applications (ICOA), 2018, pp. 1-4

[7] A. Lakshmanarao, A. Srisaila and T. S. R. Kiran, "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), vol. 23, no. 2, pp. 980-988, 2021

[8] V. Gupta, V. Aggarwal, S. Gupta, N. Sharma, K. Sharma and N. Sharma, "Visualization and Prediction of Heart Diseases Using Data Science Framework," Second International Conference on Electronics and Sustainable Communication Systems (ICESC), vol. 1, no. 2, pp. 1199-1206, 2021

[9] T. Santhanam and E. P. Ephzibah, “Heart Disease Prediction Using Hybrid Genetic Fuzzy Model”, Indian Journal of Science and Technology, vol. 8, no. 23, pp: 797–803, 2015

[10] G. Purusothaman and P. Krishnakumari, “A Survey of Data Mining Techniques on Risk Prediction: Heart

Disease”, Indian Journal of Science and Technology, vol. 12, no. 3, pp. 124-131, 2015

[11] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," International Conference on Electronics and Sustainable Communication Systems (ICESC), vol. 1, no. 56, pp. 758–766, 2020

[12] V. Sharma, A. Rasool and G. Hajela, “Prediction of Heart disease using DNN”, Second International Conference on Inventive Research in Computing Applications (ICIRCA), vol. 10, no. 7, pp. 554-562, 2020

[13] Victor Chang, Vallabhanent Rupa Bhavani, MA Hossain, “An artificial intelligence model for heart disease detection using machine learning algorithms”, Journal of Healthcare Analytics, vol. 2, no. 5, pp. 342-350, 2022

[14] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021

[15] H. D. Park, Y. Han and J. H. Choi, "Frequency-Aware Attention based LSTM Networks for Cardiovascular Disease," 2018 International Conference on Information and Communication Technology Convergence (ICTC), 2018, pp. 1503-1505

[16] Y. An, N. Huang, X. Chen, F. Wu and J. Wang, "High-Risk Prediction of Cardiovascular Diseases via Attention-Based Deep Neural Networks," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 3, pp. 1093-1105, 1 May-June 2021

[17] W. Zeng, X. Wang, K. Xu, Y. Zhang and H. Fu, "Prediction of cardiovascular disease survival based on artificial neural network," 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), 2021, pp. 219-224

[18] D. M. R., S. Kuwelkar and R. Sivakumar, “An hybrid technique for optimized clustering of EHR using binary particle swarm and constrained optimization for better performance in prediction of cardiovascular diseases”, Measurement: Sensors, vol. 9, no. 4, pp. 1376-1390, 17 December 2022

[19] K. Junwei, H. Yang, L. Junjiang and Y. Zhijun, "Dynamic prediction of cardiovascular disease using improved LSTM," in International Journal of Crowd Science, vol. 3, no. 1, pp. 14-25, April 2019

[20] V. S. Dehnavi and M. Shafiee, "The risk prediction of heart disease by using neuro-fuzzy and improved GOA," 2020 11th International Conference on Information and Knowledge Technology (IKT), 2020, pp. 127-131

[21] P. Theerthagiri, “Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique”, Intelligent Systems with Applications, vol. 8, no. 5, pp. 1079-1094, 6 September 2022

[22] R. Li, S. Yang and W. Xie, "Cardiovascular Disease Prediction Model Based on Logistic Regression and Euclidean Distance," 2021 4th International Conferenc on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2021, pp. 711-715

[23] A. Elbadry and S. Eldawlatly, "Majority-Vote Over Multiple ECG Segments for Risk Assessment (MOMESRA): A Machine Learning Approach for Predicting Cardiovascular Events," 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), 2021, pp. 1-6

[24] N. S. Rajjliwal and G. Chetty, "Deep Learning Based Decision Support Framework for Cardiovascular Disease Prediction," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2021, pp. 1-1