Prediction Techniques of Heart Disease and Diabetes Disease using Machine Learning

Dr.Geetha.S[a], Dr.Punitha Devi. C[a], Kalaivani. V[a], Haritha.C.J[a], and Preetha.G[a]

[1] Department of Information Technology, Sri Manakula Vinayagar Engineering College, Puducherry.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Heart disease and Diabetes disease is one of the most common diseases. These diseases are quite common nowadays so we used different attributes which can relate to these diseases well to find the better method to predict and we also used algorithms for prediction. Generally, People in IT sectors are becoming stressed due to their busy schedules and targets. So, they don't have sufficient time to take care of their health and families. To overcome this, we have created a website named MEDCARE to collect the sensor data and to produce the result. Notwithstanding this weight is the serious issue which is making a significant effect in everybody's life. So that in this web application they can likewise see their wellbeing status by weight list (Body Mass Index). Random Forest Classifier and K Nearest Neighbour, algorithm is analyzed on data set based on risk factors. Here the trained data sets and incoming test cases are processed by a machine learning algorithm and produce the results accordingly. Perform enlightening examination on heart disease illness forecast, bosom malignancy expectation and diabetes forecast utilizing key components like Glucose levels, Blood Pressure, Skin Thickness, BMI and so forth Outwardly investigate these factors, you may have to search for the dissemination of these factors utilizing histograms. On the off chance that they neglect to screen their health status the application will inform the employee to deal with their health.

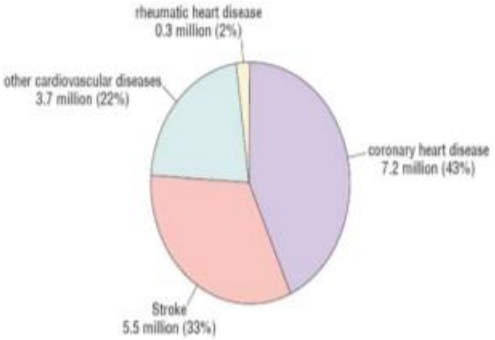
Keywords: Machine Learning, Prediction, Histograms, Diabetes forecast

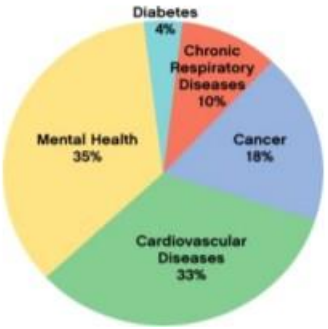
1. Introduction

Our current lifestyle is getting more pleasing, and especially imaginative. Close to the comfort that development has brought, it moreover propels bothersome lifestyle affinities, for instance, negative dietary examples, less exercise and lack of sleep. These affinities can unfavorably influence a person's life on account of their heavy stress in work. Stress depicts the strain level brought about by everyday requests just as numerous different factors, for example, family, work, and social issues. For the most part in IT areas individuals are getting more focused by considering their consumption of undertakings and targets with the goal that they principally neglected to deal with their wellbeing.

Because of this by not dealing with their wellbeing, now and then it has become a significant issue and will prompt passing in little ages. Heart disease and diabetes disease is the kind of disease which can cause death. Heart disease can be detected using the symptoms like: high blood pressure, chest pain, hypertension, cardiac arrest, etc. For diabetes often, there are no symptoms. When symptoms do occur, they include excessive thirst or urination, fatigue, weight loss or blurred vision. Nowadays there are too many automated techniques to detect heart disease like data mining, machine learning, deep learning, etc. So, in this paper we will briefly introduce machine learning techniques. In this we train the datasets using the machine learning repositories. There are some risk factors on the basis ofthat the heart disease is predicted. The main topic is prediction using machine learning techniques.

Machine learning is widely used nowadays in many business applications like e commerce and many more. Prediction is one of the areas where this machine learning is used, our topic is about prediction of heart disease and diabetes disease by processing a patient's dataset and a data of patients to whom we need to predict the chance of occurrence of those diseases.





Heart disease can be detected using the symptoms like: high blood pressure, chest pain, hypertension, cardiac arrest, etc. There are many types of heart diseases with different types of symptoms. Like: 1) heart disease in blood vessels: chest pain, shortness of breath, pain in neck throat., 2)heart disease caused by abnormal heartbeats

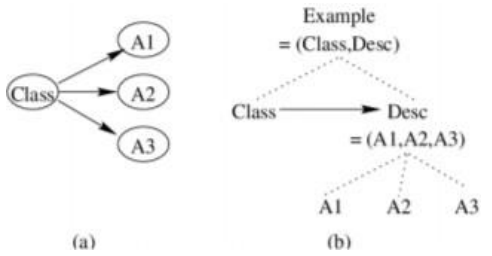
:slow heartbeat, discomfort, chest pain., etc. Most common symptoms are chest pain, shortness of breath, discomfort, chest pain., etc. Most common symptoms are chest pain, shortness of breath, fainting. Causes of heart disease are defects you're born with, high blood pressure, diabetes, smoking, drugs, alcohol. Sometimes in heart disease the infection also affects the inner membrane which is identified by symptoms like fever, fatigue, dry cough, skin rashes. Causes of heart infection are bacteria, viruses, parasites. Types of heart disease: Cardiac arrest, Hypertension, Coronary artery disease, Heart failure, Heart infection, Congenital heart disease, Slow heartbeat, Stroke type heart disease, angina pectoris. Nowadays there are too many automated techniques to detect heart disease like data mining, machine learning, deep learning, etc. So, in this paper we will briefly introduce machine learning techniques. In this we train the datasets using the machine learning repositories. There are some risk factors on the basis of that the heart disease is predicted. Risk factors are: Age, Sex, Blood pressure, Cholesterol level, Family history of coronary illness, Diabetes, Smoking, Alcohol, Being overweight, Heart rate, Chest Pain.

2.Literature Review

Literature review presents the existing survey on a particular topic which may act as the prop for the proposed systems.

2.1 Heart disease Prediction using Naïve Bayes algorithm:

In this the algorithm used was Naive Bayes algorithm. In Naïve Bayes algorithm they used Bayes theorem. Hence Naive Bayes has a very power to make assumptions independently. The used data-set is obtained from a diabetic research institute of Chennai, Tamilnadu which is a leading institute. There are more than 500 patients in the dataset. The tool used is Weka and classification are executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%.

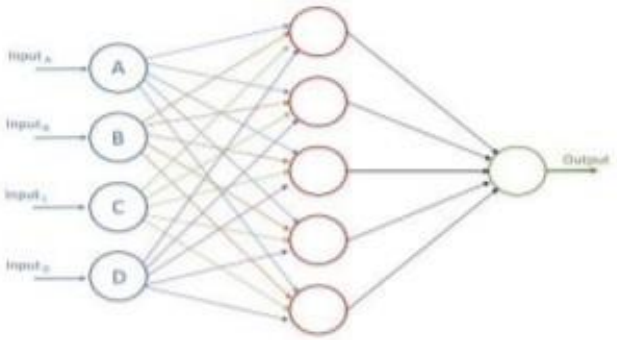


Limitation:

It is not a good way to describe the heart disease for all parameters, here in this the algorithm performs poorly for the numeric data.

2.2**Heart Disease Prediction using Neural Network Algorithm:Proposed Model:**

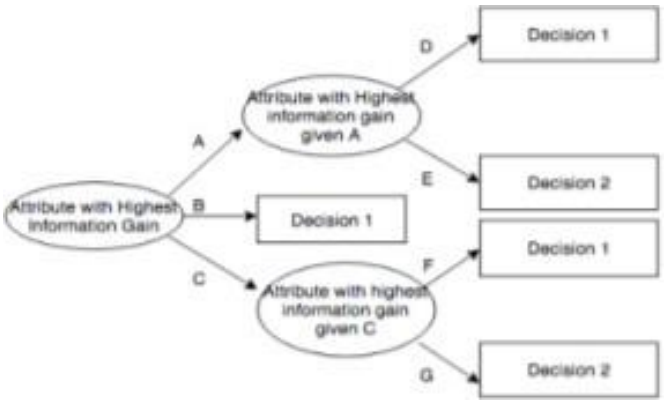
In this paper proposed system they used the neural network algorithm multi-layer perceptron (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in the input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which is added to the node to balance the perceptron. The connection between the nodes can be feedforward or feedback based on the requirement.



Limitations
The major drawback of this model is it takes more time for processing and the given accuracy rate is very low.
Unexplained working of the network.

2.3 Heart Disease Prediction using ID3 Algorithm: Proposed Model:

In this paper prediction for similarities of disease by using ID3 algorithm in television and mobile phone. This paper gives a programmed and concealed way to deal with recognized designs that are covered up of coronary illness. The given framework utilizes information mining methods, for example, ID3 algorithm. This proposed method helps the people not only to know about the diseases but it can also help to reduce the death rate and count of disease affected people.

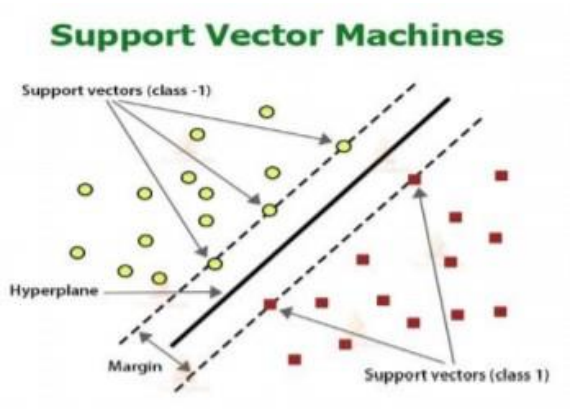


Limitations:

The main drawback is only one attribute at a time is tested for making a decision.
Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

2.4 Heart Disease Prediction using Support Vector Machine Algorithm:Proposed Model:

In this paper prediction Support Vector Machine is an extremely popular supervised machine learning technique (having a predefined target variable) which can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall.

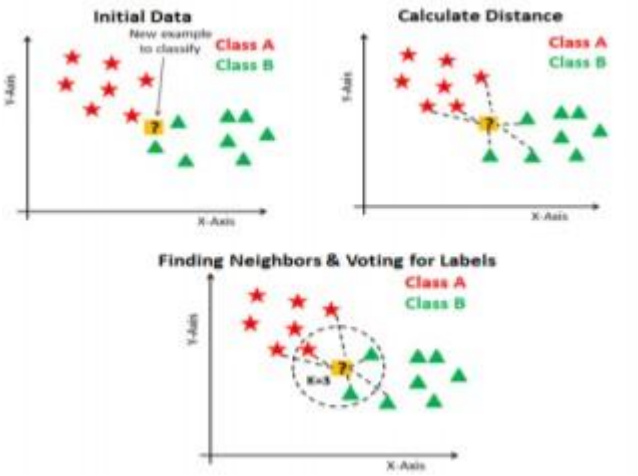


Limitations:
The main drawback is SVM algorithm not suitable for large data sets.
In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.

2.5 Heart Disease Prediction using K Nearest Neighbor Algorithm:Proposed Model:

In this paper prediction is done using KNN. K-means clustering is one of clustering techniques used to cluster

datasets based on nearest-neighbor. Here the data is clustered in k clusters based on a similarity between them. We also fill missing values of data using this k-means. Once we clustered the data every dataset will come into any one of clusters by using this clusters if we have missing values in dataset, we can fill those values as this are categorized into groups. Now as these missing values are all cleared, we can apply different prediction techniques on this for an example we can apply now as we know that for a dataset to be used for prediction in Naïve Bayes need to be pre-processed. By using this algorithm we can achieve good accuracy.



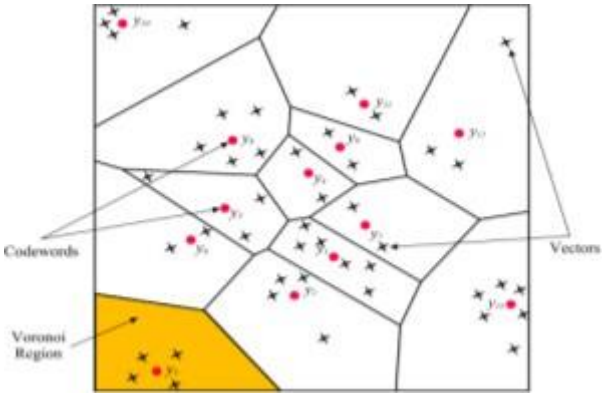
Limitations:

The main drawback is KNN algorithm needs high memory.

With large data, the prediction stage might be slow and produces low accuracy rate.

2.6 Heart Disease Prediction Using Quantization Algorithm:Proposed Model

In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation The neural system in this framework acknowledges 13 clinical includes as information and predicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures.



Limitations:

The main drawback is that as vectors of full length are used, at higher bit-rates the computational complexity and memory requirements increases in an exponential manner making it impractical for applications requiring higher bit-rates.

3.Existing System

In this paper, a comparative analysis of different classifiers was performed for the classification of the Heart Disease dataset in order to correctly classify and or predict HD cases with minimal attributes.

The set contains 76 attributes including the class attribute, for 1025 patients collected from Cleveland, Hungary,

Switzerland, and Long Beach, but in this paper, only a subset of 14 attributes is used, and each attribute has a given set value. The algorithms used K- Nearest Neighbor (K-NN), Naive Bayes, Decision tree J48, JRip, SVM, Adaboost, Stochastic Gradient Descent (SGD) and Decision Table (DT) classifiers to show the performance of the selected classifications algorithms to best classify, and or predict, the HD cases.

4. Result of Existing System:

It was shown that using different classification algorithms for the classification of the HD dataset gives very promising results in term of the classification accuracy for the K-NN (K = 1), Decision tree J48 and JRip classifiers with accuracy of classification of 99.7073, 98.0488 and 97.2683% respectively. A feature extraction method was performed using Classifier Subset Evaluator on the HD dataset, and results show enhanced performance in term of the classification accuracy for K-NN (N = 1) and Decision Table classifiers to 100 and 93.8537% respectively after using the selected features by only applying a combination of up to 4 attributes instead of 13 attributes for the prediction of the HD cases.

4.1 Parameter's Sensitivity:

The training sample size for Naive Bayes classifier will be used as a sensitivity parameter, by changing its training set size and observing the changes in its classification accuracy with respect to the portion of the training samples with respect to the total samples. Naïve Bayes was selected as an example of low accuracy rate classifier, and to see the changes of its performance in terms of the changes of the training sample size. Regarding the sensitivity analysis, parameters start with the default value of the parameter, then it was changed accordingly to study the changes of the classifier performance in terms of these parameters.

Drawbacks of Existing System:
The existing system which has been design have a complicated architecture. Most of the system which had been developed doesn't process on a large dataset.
Detection is not possible at an earlier stage.
In the existing system, practical use of collected various data is time consuming.
Accuracy is one of the major drawbacks while looking into the existing system the prediction could not be done accurately.

5 Proposed Method

We have created a web application for 2 diseases and can view their BMI health status.

Heart disease

Diabetes disease

BMI (Body Mass Index)

The diseases are predicted using machine learning algorithms such as Random Forest is used for heart disease and K Nearest Neighbor is for Diabetes.

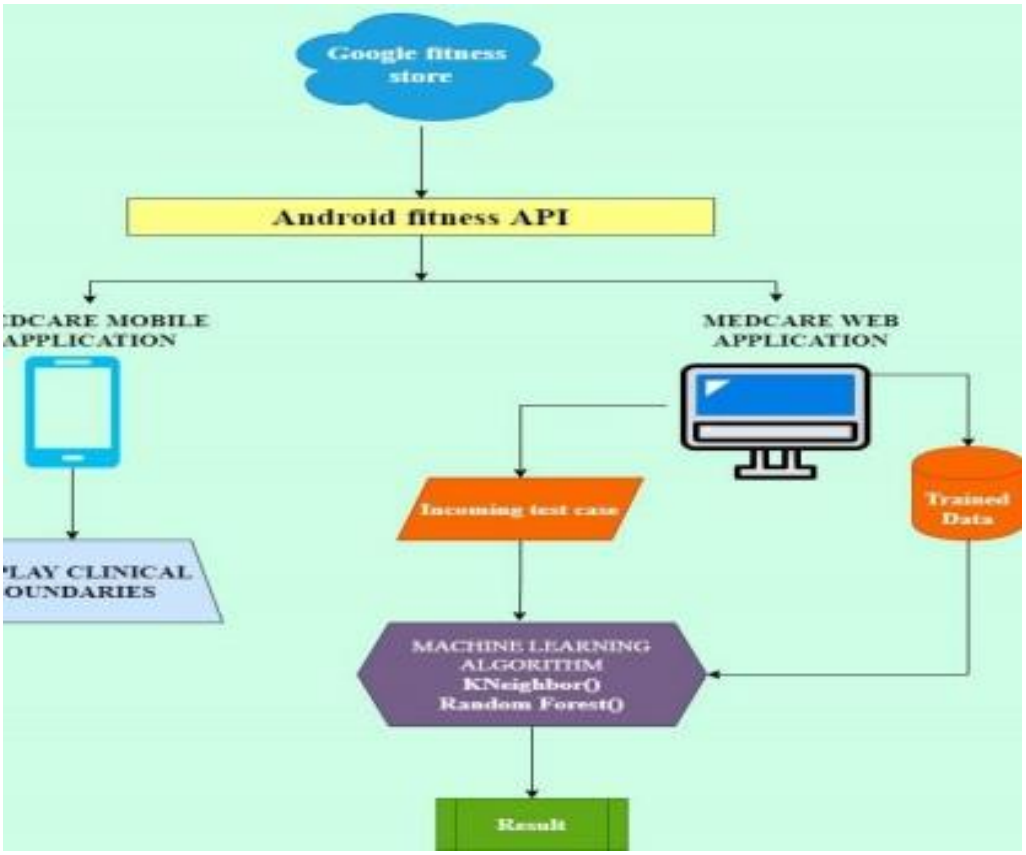


Fig 3: Work flow design

5.1 Prediction Analysis:

Perform descriptive analysis on heart disease prediction, and diabetes prediction using key factors like Glucose levels, Blood Pressure, Skin Thickness, BMI etc. Visually explore these variables, you may need to look for the distribution of these variables using histograms. Treat the missing values accordingly. We observe integers as well as float data-type of variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables. Check the balance of the data byplotting the count of outcomes by their value. Describe your findings and plan future course of actions.

Create scatter charts between the pair of variables to understand the relationships. Describe yourfindings. Perform correlation analysis. Visually explore it using a heat map. Devise strategies for model building. It is important to decide the right validation framework.

Apply an appropriate classification algorithm to build a model. Compare various models with the results from Random Forest Algorithm and K Nearest Neighbor.

5.2 Heart Disease Prediction Analysis:

There are 768 rows and 9 columns in this dataset.
The dataset has various attributes like Sex, age, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope.
While the maximum for age reaches 77, the maximum of chol (serum cholesterol) is 564.

Diabetes Prediction Analysis:

There are 768 rows and 9 columns in this dataset.
The dataset has nine attributes in which there are eight independent variables (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age) and one dependent variable (Outcome).
BMI and DiabetesPedigreeFunction are a float data type and rest of the variables are integer data type.
The Variables have a lot of zero values which can be represented as missing values.

The missing values '0' is replaced by mean to explore the dataset.

5.3 Features of Proposed System:

Machine Learning

Machine Learning is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of Machine Learning is prediction - and predictive Machine Learning is the most common type of Machine Learning and one that has the most direct business applications. The process of Machine Learning consists of three stages: (1)the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).
Machine Learning commonly involves four classes of tasks.

- ***Classification*** - Arranges the data into predefined groups. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include Yolo Classification Learning, and Neural network.

- ***Clustering*** - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.

- ***Regression*** - Attempts to find a function which models the data with the least error.

- ***Association rule learning*** - Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as marketbasket analysis.

PYTHON:
PYTHON is a popular suite of machine learning software for data analysis and predictive modeling. The original non-Java version of PYTHON was a TCL/TK back-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and make a file-based system for running machine learning experiments. The main strengths of PYTHON are that it is:

- freely available under the GNU General Public License,

- very portable because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform,

- contains a comprehensive collection of data preprocessing and modeling techniques,
- Is easy to use by a novice due to the graphical user interfaces it contains.

Python supports several standard Machine Learning tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. The Explorer interface has several panels that give access to the main components of the workbench. The Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called filtering algorithm. The Classify panel enables the user to apply classification and regressionalgorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself.
CSV File Format:

CSV is a Comma-separated values file, in which the data will be stored in a table format.CSV can be used with many spreadsheet programs like Microsoft Excel or Google Spreadsheets. It is totally different from other spreadsheet file types because you have only a single sheet in a file, and you cannot save any formulas in this format. It is a way of organizing data in a file by splitting each row into blocks.

They also serve two other primary business functions:

- ➢ CSV files are plain-text files, making them easier for the website developer to create.
- ➢ Since they're plain text, they're easier to import into a spreadsheet or another storage database, regardless of the specific software you're using
- ➢ To better organize large amounts of data

Random Forest Classifier:

Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Simply, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Simply, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

K Nearest Neighbor:

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithms. KNN is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g., linear-separability, uniform distribution, etc. The KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g., Euclidean or Manhattan etc. It then selects the K-nearest data points, where K can be any integer. Finally, it assigns the data point to the class to which the majority of the K data points belong.

6. Conclusion

We have summed up various kinds of machine learning algorithms for forecasts of coronary illness and diabetes infection. We explained different algorithms and pursued tracking down the best calculation by examining their highlights. Each calculation has given distinctive outcomes in various circumstances. Further it is examined that minimal exactness is accomplished for prescient models of coronary illness, diabetes and thus more mind boggling models are expected to build the precision of foreseeing the early those sickness. In future we will add more forecasts of sickness with high precision and least expense and complexity.By considering those information they can manage their wellbeing and can be prepared to work in concordance and peace. If they disregard to screen their wellbeing status the application will advise the worker to manage their wellbeing.

Acknowledgments
Research reported in this study was supported by The Department of Information Technology and The Management, Sri Manakula Vinayagar Engineering College, Madagadipet, Pondicherry.

References

1. Abhay Kishore1, Ajay Kumar2, Karan Singh3, Maninder Punia4, Yogita Hambir5," Heart Attack Prediction Using Deep Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04 | Apr-2018.
2. Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD)," Prediction of Heart Disease Using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1- 5386-0965-1.
3. A.Lakshmana Rao, Y.Swathi, P.Sri Sai Sundareswarar," Machine Learning Techniques For Heart Disease Prediction", International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019.
4. Avinash Golande, Pavan Kumar T," Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
5. Ezhilarasi, G.Dilip, T.P.Latchoumi, K.Balamurugan* (2020), UIP—A Smart Web Application to Manage Network Environments, Advances in Intelligent systems and computing book series, https://doi.org/10.1007/978-981-15-1480-7_8, 97-108.
6. Latchoumi T. P, K. Balamurugan, K. Dinesh and T. P. Ezhilarasi, (2019). Particle swarm optimization approach for water-jet cavitation preening. Measurement, Elsevier, 141,184-189.
7. Latchoumi T. P, T. P. Ezhilarasi, K. Balamurugan (2019), Bio-inspired Weighed Quantum Particle Swarm Optimization and Smooth Support Vector Machine ensembles for identification of abnormalities in medical data. SN Applied Sciences (WoS), 1137, 1-12, DOI: 10.1007/s42452-019-1179-8.
8. Latchoumi, T. P., Reddy, M. S., & Balamurugan (2020), K. Applied Machine Learning Predictive Analytics to SQL Injection Attack Detection and Prevention. European Journal of Molecular & Clinical Medicine, 7(02), 3543-3553
9. M. S. Amin, Y. K. Chiam, K. D. Varathan,''Identification of significant features and data mining techniques in predicting heart disease,'' Telematics Inform., vol. 36, pp. 82–93, Mar.2019.
10. N. Al-milli, Backpropagation neural network for prediction of heart disease, ''J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp.131–135, 2013. [14] A. S. Abdullah and R. R. Rajalaxmi, ''A data mining model for predicting coronary heart disease using random forest classifier,'' in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.
11. Pruthviraju G, K.Balamurugan*, T.P.Latchoumi, Ramakrishna M (2021), A Cluster-Profile Comparative Study on Machining AlSi7/63% of SiC hybrid composite using Agglomerative Hierarchical Clustering and K-Means, Silicon, 13, 961–972, DOI: 10.1007/s12633-020-00447-9, Springer.
12. Santhana Krishnan.J, Dr.Geetha.S," Prediction of Heart Disease Using Machine Learning Algorithms",2019 1st International Conference on Innovations in Information and Communication Technology(ICIICT),doi:10.1109/ICIICT1.2019.87414 65.
13. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava ―Effective Heart Disease Prediction Using Hybrid Machine LearningTechniques‖, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019 S.P. Bingulac, ―On the Compatibility of Adaptive Controllers,‖ Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994. (Conference proceedings).
14. S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain,''Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis,'' Phys. A, Stat. Mech. Appl.,vol. 482, pp. 796–807,2017. doi:10.1016/j.physa.2017.04.113.
15. Sonam Nikhar, A.M. Karandikar" Prediction of Heart Disease Using Machine Learning Algorithms" International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication,[Vol-2, Issue-6, June- 2016].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.
16. Stephen F. Weng, Jenna Reps, Joe Kai1, Jonathan M. Garibaldi, Nadeem Qureshi,―Can machine-learning improve cardiovascular risk prediction using routine clinical data?‖, PLOS ONE | https://doi.org/10.1371/journal.pone. 0174944 April 4, 2017.
17. V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods", International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013.
18. Vijay Vasanth A,Latchoumi T.P, Balamurugan Karnan,Yookesh T.L (2020) Improving the Energy Efficiency in MANET using Learning-based Routing, Revue d'Intelligence Artificielle, 34(3), pp 337-343.
19. Venkata Pavan M,Balamurugan Karnan*, Latchoumi T.P (2021), PLA-Cu reinforced composite filament:

Preparation and flexural property printed at different machining conditions, Advanced Composite Materials, https://doi.org/10.1080/09243046.2021.1918608

20. V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja," Heart disease prediction using machine learning techniques: a survey", International Journal of Engineering & Technology, 7 (2.8) (2018) 684-687.