Machine Learning Techniques Usage in Prediction of Multiple Chronical Diseases

[1]A. Bhargavi, [2]M. Jagadeesh Kumar, [3]B. Kishor Kumar, [4]M. Mary Sujatha

[1,2,3]P.G Students, [4]Assistant Professor
Department of Computer Science
National Sanskrit University
Tirupati- 517507, Andhra Pradesh.

Abstract- **Multiple Chronical Disease Prediction System is an effective healthcare predictive application that aims to predict multiple chronical diseases including Chronical Kidney Disease, Heart Disease, Diabetes, Brain Stroke and Lung Cancer with the help of various machine learning algorithms. Scope of the project is all-inclusive, focussing to predict the possibility of various diseases in the human beings taking into account their distinctive health profiles. The electronic health records are acquired from the online sources like UCI machine learning repository, GitHub and Kaggle. This prediction system uses data mining techniques for completing the purpose i.e., data pre-processing and is trained on various ensemble methods like random forest, supervised techniques, unsupervised techniques like decision tree, logistic regression, multi-layer perceptron and naive bayes. As most of these diseases share some common risk factors through our work, we are trying to explore the possible interconnection between these chronical diseases and also the chance of developing these chronical illnesses. Finally, this prediction system with the power of machine learning techniques helps in the identification and prognosis of such diseases at early stages to prevent the extremity of them and at the same time reducing the health care expenditure.**

Keywords: **Machine Learning Algorithms, Data Mining Techniques, Random oversampling, Chronical Disease Prediction.**

1.Introduction

Artificial intelligence (AI) is a field that encompasses machine learning (ML), enabling computers to acquire knowledge autonomously without explicit programming. ML models are utilized to render forecasts or determinations based on data. The implementation of ML methodologies for anticipating various chronical illnesses has garnered significant attention in recent times [1]. This interest stems from several factors. To begin with, chronical ailments pose a substantial public health challenge, standing as the primary cause of mortality and incapacity on a global scale. Furthermore, diagnosing and treating chronical diseases is often intricate due to their multifaceted origins in genetics, surroundings, and lifestyle choices. Additionally, there is an escalating demand for early detection and prompt intervention regarding chronical diseases to enhance patient outcomes and curtail healthcare expenses [2]. The application of ML strategies harbours the potential to enhance prognostication regarding multiple chronical maladies. By scrutinizing extensive datasets, ML models can discern trends, spot regularities, and deliver prognostications. This holds the promise of refining the timely identification and diagnosis of chronical conditions. Furthermore, ML models can aid in crafting individualized therapy blueprints for individuals grappling with chronical ailments.

Despite the promise that ML methodologies hold, there exist hurdles in integrating them for the prognostication of multiple chronical diseases [3]. Firstly, the data employed to train ML models is frequently deficient or imprecise, engendering potentially erroneous prognostications. Secondly, ML models have the propensity to exhibit bias, potentially leading to unfair or prejudiced predictions. Lastly, ML models can be challenging to decipher, making it arduous to discern the rationale behind a specific prognosis [4]. Notwithstanding these challenges, ML methodologies offer the potential to advance the prognostication of diverse chronical diseases. By harnessing ML algorithms to sift through copious datasets, recognize trends, and generate forecasts, the early identification and diagnosis of chronical ailments can be markedly enhanced [5]. Furthermore, ML models can be instrumental in formulating tailored treatment schemes for individuals afflicted with chronical diseases.

2. Preliminaries

2.1. Chronic Diseases

According to National Centre for Chronical Disease Prevention and Health Promotion (NCCDPHP) chronical diseases which are also termed as non-communicable diseases are the type of illnesses that last for a very long period lasting several years. These diseases can only be controlled but are neither cured nor prevented by medicines and

vaccines. The major cause of these diseases is tobacco usage, second hand smoke exposure, poor nutrition, excessive alcohol use, unhealthy food habits and physical inactivity. These diseases can commonly develop along with the gradual increase in age. Some of the common chronical health conditions are diabetes, heart disease, chronical kidney disease, stroke, cancer and chronical lung disease [6]. Cardiovascular diseases like brain stroke and heart disease which are chronical in nature, are the two major conditions accounting for the enormous number of deaths. These diseases are caused mainly because of tobacco, lack of nutritious food, and very less physical activity. All these activities can be managed by maintaining a healthy lifestyle. When the risk behaviours are minimized the need to predict control and prediction of the illness also decreases [7]. Following these cardiovascular diseases there are lung cancer and breast cancer which are considered as the some of the deadliest diseases. One of the major chronical illnesses is chronical kidney disease. Chronical kidney disease has varied stages of seriousness as it gets only worse over the time [8]. Diabetes is another serious and chronical health condition.

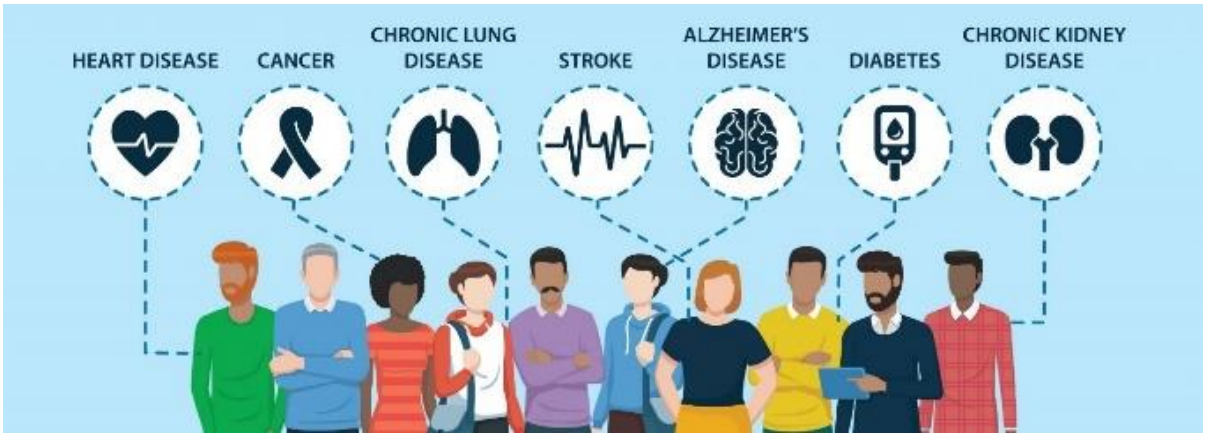

Figure 1. Chronic Diseases Management adapted from google images

People who are diabetic are at a higher risk for developing cardiovascular conditions like heart disease, brain stroke and other major complications like chronical kidney illness and amputation of body parts like toes, feet and legs. Elder people who are at the age of 45 or above are affected by this condition more specifically by type 2 diabetes [9].

3.Methodology

This segment mainly focusses on the detailed description of how the data set is created, preparation of the model, and how the disease prediction works have been discussed. The first step of this whole process involves data collection. The proposed system collects structured data which is obtained from various sources in online mode. The data is splitted into training and testing data sets. The model is trained on training data set with the different machine learning algorithms such as logistic regression, random forest classifier, decision tree classifier, linear and non-linear support vector machines, two boosting classifiers, and gaussian naive bayes [10]. The next measure is to test the model with testing data for finding out the algorithm which has the best accuracy amongst the all. That model is then tested to evaluate its performance with new data that is not used for training purpose. If the model obtains the desired accuracy in testing data, then the developed model can be used for deployment.

3.1. Data Collection

The collection of data sets for the five chronical diseases, including chronical Kidney disease, Heart disease, Diabetes, Brain stroke, and Lung cancer are taken from multiple online sources like UCI machine learning repository, GitHub and Kaggle.

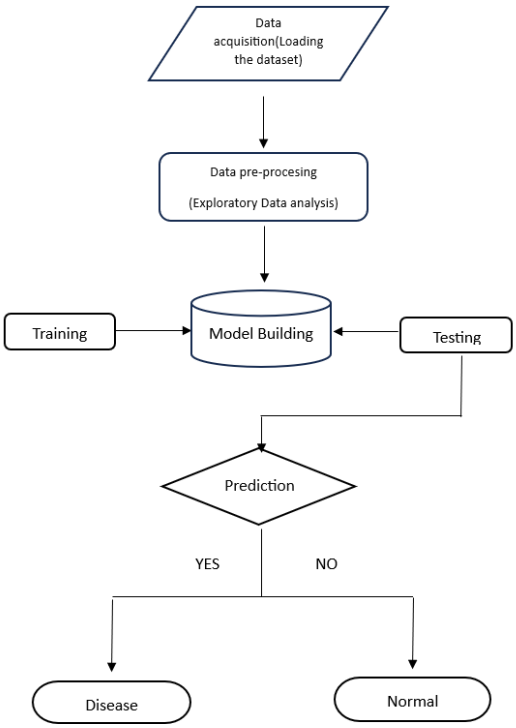

Fig 2. Workflow of the model

All datasets acquired consists the existing medical conditions which are essential, life style habits. Personal information of the patient like ID number, number of visits to the hospital are excluded to secure the patient's privacy. and other details like marriage status, employment status, type of residency which are irrelevant are excluded and not considered for the data pre-processing step. Most features of the collected datasets are of categorial, binary and integer types. Kidney disease dataset contains a total of 401 records of 25 attributes. Some of them are patient id, age, gender, albumin, blood pressure, sugar, pus cell, blood glucose random level bacteria, blood urea, serum creatinine, sodium levels, potassium levels, haemoglobin, packed cell volume, hypertension, coronary artery disease, diabetes, red blood cells count, white blood cells count, appetite, pedal edema, anemia [11]. Diabetes dataset contains a total of 768 records of 8 significant attributes and they are blood pressure, number of pregnancies , thickness of the skin, insulin levels, diabetes pedigree function, bmi of the patient .Heart disease dataset contains 13 attributes and some of them are age, gender, chest pain type experienced, blood sugar levels, electrocardiogram results, cholesterol level, exercise induced angina etc., Stroke prediction dataset contains 11 attributes like gender, age of the person, hypertension, glucose level, existing heart disease, marital status, type of work, type of residency, bmi level of the patient, smoking habit. The model implementation starts with loading the acquired the data and loading the dataset. Then the data is pre-processed i.e., data cleaning is done. Now the model undergoes training and testing to measure the accuracy achieved by the it. Finally, the model with highest accuracy is used for disease prediction.

3.2. Data Pre-Processing

Data pre-processing is a standard artificial intelligence technique to transform the available raw data into useful information. In predictive analysis the quality of data is a crucial factor, low quality of the data leads to incorrect results. So, the acquired data is pre-processed to fill out the missing values, removal of the outliers and irrelevant features in most of the structured data. This makes the machine learning algorithm more productive and efficient. The data pre-processing is essential before using data for training the model. After the preprocessing of data has been completed, then it can be used for prediction purpose.

3.3. Model Building and Training

After the completion of data pre-processing the dataset is now divided into two classes i.e., 'x' and 'y' where x is the feature columns and y is the target label. Next these x and y classes are splitted into training and testing data with a format of 70:30 using train _test split method. 70% of data is used for training the model and 30% of the data is used for testing and validating the model. Following the splitting of data, several machine learning models are fitted to the pre-processed data and the accuracy of each and every model is measured. Following are the machine learning techniques applied for model building and training.

Random Forest Classifier can surely be named as a popular supervised technique of machine learning which is applicable to both regression and classification issues in machine learning. The base concept the random forest algorithm is ensemble learning method. The base estimator of this ensemble technique is decision trees [12]. The algorithm predicts the output considering the outputs of the various number of decision trees present in the forest. Gradient

Boosting Classifier is a supervised technique of machine learning. The fixed base estimator for this ensemble technique are decision trees. Along with base estimator it aggregates several other weaker learning models to form a strong and final learning model. It reduces the errors in prediction by considering the number of errors occurred in the previous iterations and at the same time optimizing the weights present in the model [13]. Decision Tree Classifier which is also a supervised tree- structured classifier is based on CART (Classification and Regression Tree) algorithm. It is a mostly preferred for solving complex problems of classifications because of their ability to mimic human thinking and also the simple logic behind the working of the tree. It uses a technique called Attribute selection measure which further enhances the model's accuracy. Multilayer Perceptron is a machine learning technique which comes under supervised learning. Multi-layer perceptron is a class of feedforward neural network in which all the nodes are fully connected. It comes under the category of artificial neural networks. By using back propagation algorithm in training, it increases the model's lower accuracy by the reducing the error rate.

XG boosting Classifier or extreme boosting is supervised machine learning technique with the base concept of ensemble method. Gradient descent which is an optimization approach used for the training purpose [14]. It uses decision tree as its base estimator. Along with the base estimator it builds a final model by combining various independent estimators for the predicting the output. K Nearest Neighbour is a simple non-parametric machine learning technique. It is also named lazy learner algorithm and is used if a very big amount of data is present. After setting the K value for number of neighbours and calculating the Euclidean distance between them it categorizes the fresh data points putting them into the available category which is most similar to it. Support Vector is a simple but a potent supervised machine learning technique which works by creating a best line of boundary decision for classifying the data points. This technique is mostly used for smaller datasets which are complex in nature [15]. As there are two types of SVM namely linear and non-linear we have used both of them for the prediction.

4. Experimental Results

After the machine learning algorithms are trained on the acquired datasets it is found that random forest classifier and decision tree classifier achieved the highest accuracy in comparison to all other algorithms. Four performance metrics which are accuracy, F1-score, recall, precision are used to assess the accuracy of proposed model. These are the important measures which are used for solving any classification problem. Among all the performance matrices accuracy is the most important one as it the only factor which determines whether the person is diseased or normal. The accuracy of all the classification algorithms is shown below:

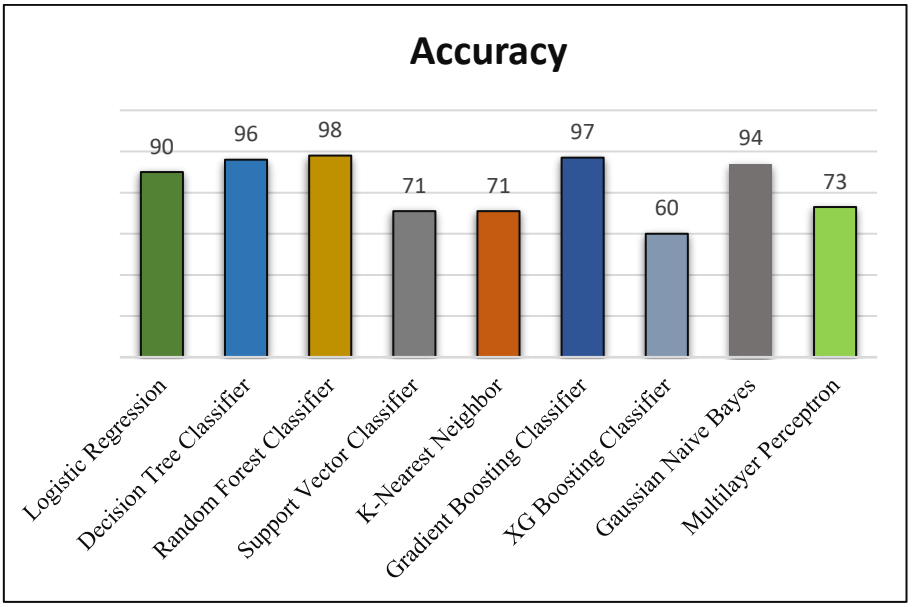

Fig 3. Chronic Kidney Disease Diagnosis

Fig 3. shows that when various machine learning models are trained and tested random forest achieved the highest accuracy 98% out of all the machine learning techniques in chronic kidney disease prediction.

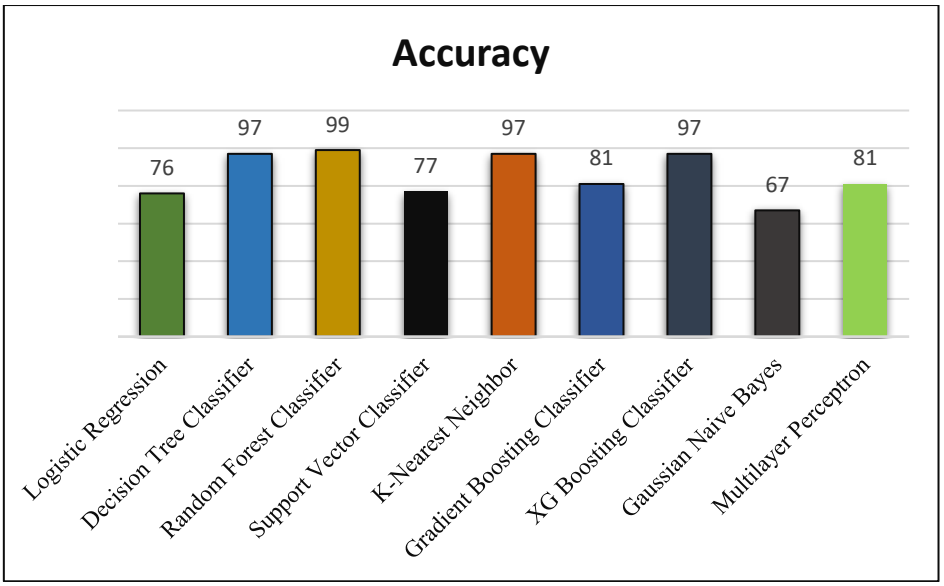

Fig 4. Stroke Diagnosis

Fig 4. shows that when various machine learning models are trained and tested random forest achieved the highest accuracy 99% out of all the machine learning techniques in the diagnosis of brain stroke.

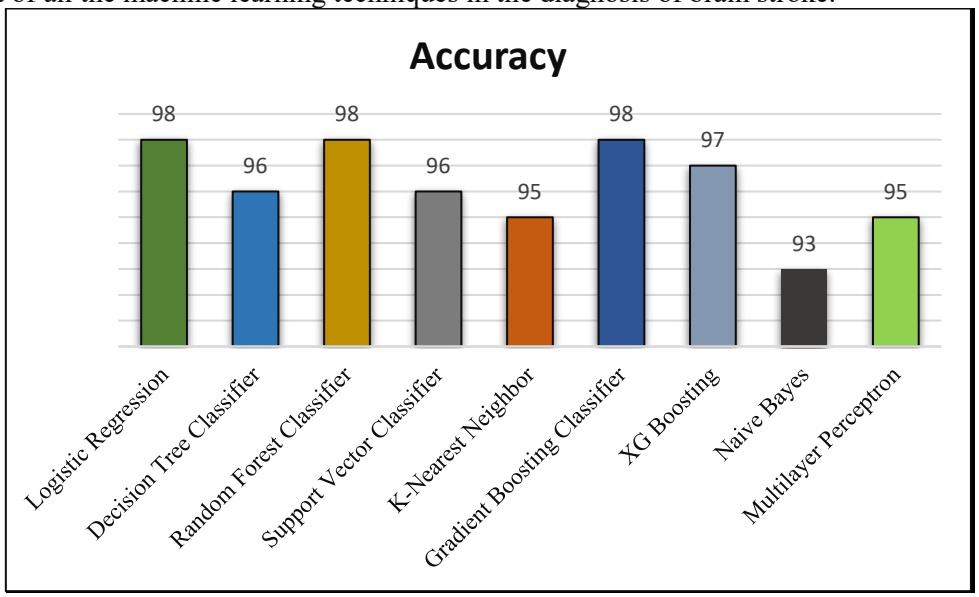

Fig 5. Lung Cancer Diagnosis

Fig 5. shows that when various machine learning models are trained and tested random forest, logistic regression and gradient boosting achieved the same highest accuracy 98% out of all the machine learning techniques in the lung cancer diagnosis.

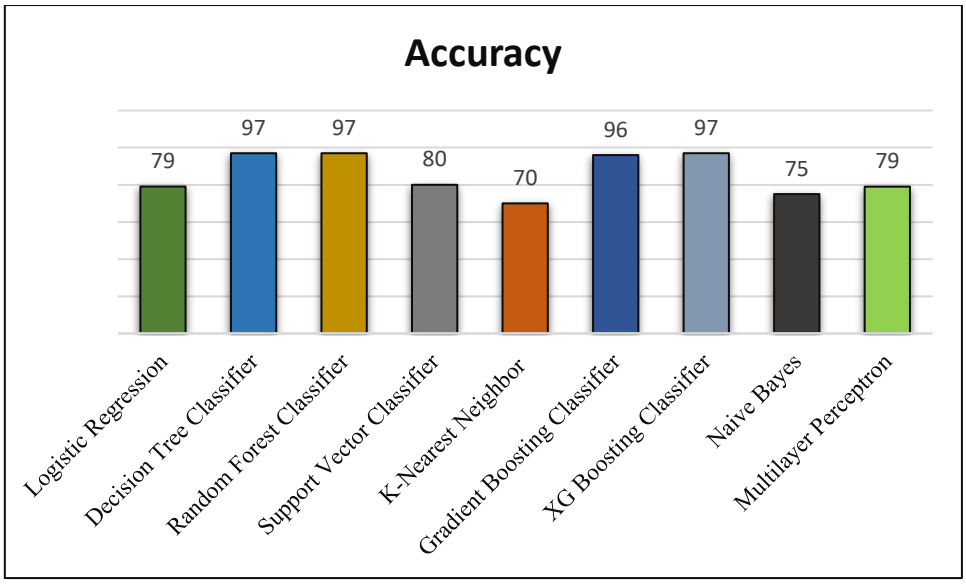

Fig 6. Heart Disease Diagnosis

Fig 6. shows that when various machine learning models are trained and tested decision tree, random forest and XG boost achieved the same highest accuracy 97% out of all the machine learning techniques in the heart disease diagnosis.

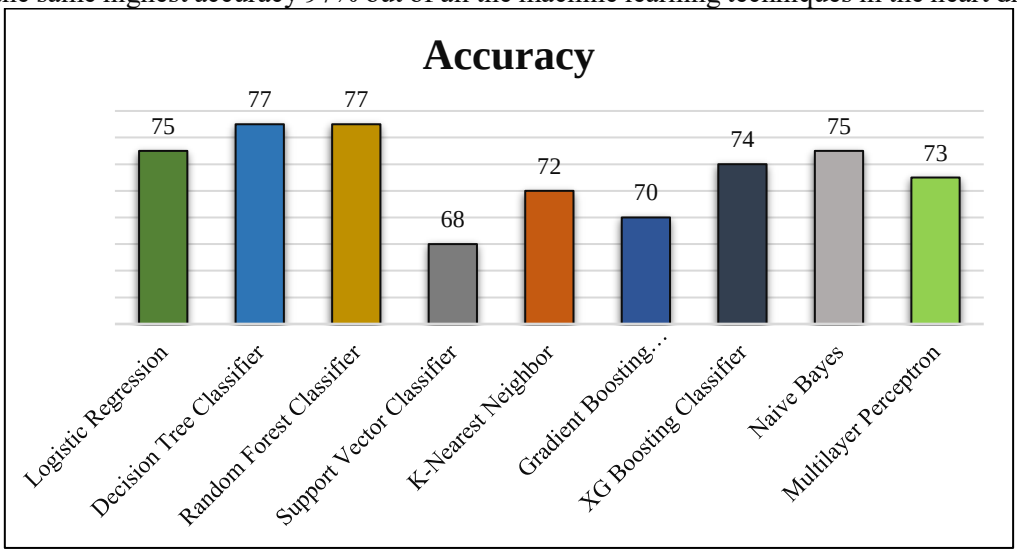

Fig 7. Diabetes Diagnosis

Fig 7. shows that when various machine learning models are trained and tested decision tree and random forest achieved the same highest accuracy 97% out of all the machine learning techniques in the diagnosis of diabetes.

Accuracy of the various machine learning approaches %									
	Logistic Regression	**Decision Tree Classifier**	**Random Forest Classifier**	**Support Vector Classifier**	**K-Nearest Neighbour**	**Gradient Boosting Classifier**	**XG Boosting Classifier**	**Naive Bayes**	**Multilayer Perceptron**
Kidney Disease	**90**	**96**	**98**	**71**	**71**	**97**	**60**	**94**	**73**
Stroke	**76**	**97**	**99**	**77**	**97**	**81**	**97**	**67**	**81**
Lung Cancer	**98**	**96**	**98**	**96**	**95**	**98**	**97**	**93**	**95**
Heart Disease	**79**	**97**	**97**	**80**	**70**	**96**	**97**	**75**	**79**
Diabetes	**75**	**77**	**77**	**68**	**72**	**70**	**74**	**75**	**73**

Table 1. Comparative analysis of all the models

Table 1. is the visual representation depicting the comparison of the accuracy results. Accuracy results of 9 different algorithms such as naive bayes, decision tree, logistic regression, k-nearest neighbour, support vector machines, gradient boosting, multilayer perceptron, xg boosting and random forest are demonstrated through the above table. Random forest, xg boosting, decision trees, and gradient boosting achieved an average of 93.8%, 92.6%, 88.4% & 85% followed by other algorithms.

5. Conclusion

Chronical conditions like heart disease, diabetes, brain stroke, chronical kidney disease and lung cancer are proven incurable and are causing death medical conditions. As these are incurable medical conditions the treatment should be started at the earliest so that the future complications can be averted. With the help of machine learning model which we have created we can easily detect these conditions in the earlier stages and thereafter minimizing their deadly outcomes. This work inspects how good various supervised and unsupervised machine learning algorithms predicts the presence of such conditions based on the person's past medical data. By scrutinizing all the machine learning techniques used it can be concluded that random forest classifier is the most suitable technique for the prediction of chronical health conditions. Supervised classifier like random forest achieved an average accuracy of 93.8% surpassing all the other algorithms. Random forest which is an ensemble technique contains more than one decision tree on different subclasses of the attributes of the dataset. The classifier considers all the predictive outputs made by each of the tree present in the forest, so this helps in improving the accuracy and making the correct predictions. The developed model can surely lower the possibility of developing the chronical health conditions by detecting them in advance which subsequently lowers the health care expenses.

REFERENCES:

[1]. Alanazi R. Identification and Prediction of Chronic Diseases Using Machine Learning Approach. *JHealth Eng*. 2022 Feb 25; 2022:2826127. Doi: 10.1155/2022/2826127. PMID: 35251563; PMCID: PMC8896926. https://doi.org/10.1155/2022/2826127

[2]. Junaid R., Saba B., Jungeun K., Muhammad W.N., Amir H., Sapna J., & Riti K. (2022). An Augmented Artificial Intelligence Approach for Chronical Disease Prediction. *Frontiers in Public Health*, Vol 10 2022,11-13 https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.860396/full

[3]. Chalapathi R T., Phaneendra V. V. S. I., Tejitha A., Sai Surya J., & Keerthi S. M. (2022). Chronical Disease Prediction Using Machine Learning. *International Journal of Scientific Research in Engineering and Management*, 6(7), 3-5, https://portal.issn.org/resource/ISSN/2582-3930

[4]. Induja S. N., Raji C. G. Computational methods for predicting chronical disease in healthcare communities. Proceedings of the 2019 *International Conference on Data Science and Communication (Icon DSC)*; March 2019; Bangalore, India. IEEE; pp. 1–6. https://ieeexplore.ieee.org/abstract/document/8817044

[5]. Rakibul Hoque Md., Sajedur Rahman M. Predictive modelling for chronical disease: machine learning approach. Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis; March 2020; Silicon Valley, CA, USA. pp. 97–101. https://dl.acm.org/doi/abs/10.1145/3388142.3388174

[6]. Hossain Md E. Camperdown NSW, Australia: University of Sydney; 2020. Predictive Modelling of the Comorbidity of Chronical Diseases: A Network and Machine Learning Approach. PhD Thesis https://ses.library.usyd.edu.au/handle/2123/24229

[7]. Rahman, S., Hasan, M., & Sarkar, A. K. (2023). Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. *European Journal of Electrical Engineering and Computer Science*, *7*(1),23–30. https://doi.org/10.24018/ejece.2023.7.1.483

[8]. Harshit J., Sarthak A., Reshab K., Rachna J., & Preeti N. (2021). Heart Disease Prediction Using Machine Learning Algorithms, *IOP Conf. Series: Materials Science and Engineering* ,1022(2021)012072, 3-7, doi:10.1088/1757-899X/1022/1/012072

[9]. Aishwarya M., & Dr. Vaidehi V. (2019). Diabetes Prediction Using Machine Learning Algorithms, *International Conference on recent trends in advanced computing,* Procedia Computer Science, 165 (2019) 292–299.

[10]. Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. *Journal of pathology informatics*, *14*,100189. https://doi.org/10.1016/j.jpi.2023.100189

[11]. E. Halim, L. Artahni, Y. Kurniawan, R. Tjahyadi and P. P. Halim, "Chronic Disease Prediction using Data Mining and Machine Learning Algorithm," 2022 *5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),* Yogyakarta, Indonesia, 2022, pp. 811-816, Doi: 10.1109/ISRITI56927.2022.10052970 https://ieeexplore.ieee.org/document/10052970

[12]. Shweta Agarwal, Dr. Chander Prabha, Dr. Meenu Gupta. (2021). Chronic Diseases Prediction Using Machine Learning – A Review. *Annals of the Romanian Society for Cell Biology*, 3495–3511. Retrieved from https://www.annalsofrscb.ro/index.php/journal/article/view/461

[13]. Pramod Reddy P, Madhu Babu D, Hardeep Kumar & Dr. Shivi Sharma (2021). Disease Prediction using machine learning. *International journal of Creative Research Thoughts*, volume 9, issue 5 May 2021,205-207

[14]. Gopi B, Getu Gamo Sagaro, Nalini C, & Francesco Amenta. (2020). Application of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized medicine*, 10(21),1-3, doi:10.3390/jpm10020021.

[15]. Hedge S. and Mundada M R. (2021), Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence-based feature selection approach, *International Journal of Pervasive Computing and Communications*, Vol. 17 No.1, pp.20-36 , https://doi.org/10.1108/IJPCC-04-2020-0018