



Hybrid Machine Learning Algorithm For Prediction Of Various Diseases.

Ravi Choubey^{1*}, Dr. Pratima Gautam²

Abstract

Now-a-days due to improper lifestyle, junk food and bad habits we are facing so many problems in our life and because of this we are inviting unwanted lethal diseases. Heart Disease, Diabetes and Hypertension are three of them and it has affected numbers of people worldwide. In recent times, These diseases have been become the major reason of death among people of any age group. Therefore, the enhancement for predicting this kind of diseases are required in the health sector with the help of different Machine Learning (ML) methods. In this study we used Hybrid Ensemble Common Model (HECM) for predicting Heart, Diabetes and Hypertension Disease. LightGBM, Random Forest, KNN used as Ensemble Classifier then output given to Voting classifier for final output. Cross Validation done at last and final output recorded .

1592

Keywords – Hybrid Machine Learning, Random Forest, Light Gradient Boosting, KNN, Disease Prediction

DOI Number: :10.14704/nq.2022.20.8.NQ44172

NeuroQuantology 2022 ;20(8):1592-1601

1. INTRODUCTION

Machine learning is an subset of Data Science that provides computers the ability to learn automatically and improve from experience without being explicitly programmed. Machine Learning focuses on the developments of system programs to access data and make prediction for future decision. Machine learning uses many Techniques and models to learn itself. While Extracting knowledge from enormous amount of data and translating unprocessed data into valuable information called Data Science . This technology provides support in the recognition of patterns

among data. There are various application fields in which data Science techniques are used extensively. These fields include many businesses , clinical diagnosis, science & engineering ,Social media etc. Now days alone Machine Learning's single classifier is not enough to classify with higher accuracy and less time . So we can ensemble many classifier to each other, this ensembling method is called Hybrid Machine Learning Model. In previous research study comparison of various classifier ensembles are used for Disease prediction. But they didn't build any common model for various diseases.

Corresponding author: Ravi Choubey

Address: 1*,2 Department of Computer Science and I.T, Rabindranath Tagore University Raisen,India.

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: XX XXXX 2022 **Accepted:** XX XXXXX 2022



In this paper we introduce Hybrid Ensemble Common Model (HECM) for various Disease prediction . during Covid-19 pandemic survival rate was low in hypertension, heart or diabetes patient and majority death in Covid-19 pandemic was due to these diseases.

1.2 Disease Description

- **Diabetes**, is concerns as a long-lasting health condition that affects the process of turning food into energy. In this condition blood sugar goes up. This disease may be lead to heart disease. Most common symptoms are urinate a lot, are very thirsty and hungry, loose weight dry skin and tired.
- **Hypertension**, that is a problem which force of the blood against the artery walls which is too high. Common symptoms are headache, irregular heartbeats , nervousness ,chest pain, blood in urine.
- **Heart attack**, when a blood clot blocks the flow of blood to the heart, and the tissues die due to lack of blood, this condition is called a heart attack. commonly seen symptoms of cardiac disease are breath shortness, physical body illness and swelling in feet.

Researchers make an attempt for discovering an effective method so that these diseases can be detected because the existing diagnosis methods of these disease are less efficient in early time identification. There are various reasons behind this including accuracy. The precise and accurate detection of above kind of illness is relied on the earlier knowledge and information regarding the pathological events.

Problem Definition

The main problem in disease prediction is to classify that the person or patient have disease or not. If patient have any disease then it classifies as unhealthy else healthy. Our problem is classifies a disease with higher accuracy and with single model.

LITERATURE REVIEW

Senthil kumar Mohan, suggested a new technique named HRFLM utilized to discover a valuable attribute with the implementation of ML schemes. The hybrid RF was integrated with the LM. This technique leads to enhance the accuracy while predicting the heart only disease. The predictive

model was presented by incorporating attributes and various classification methods. An improved performance level with only 88.7% accuracy was obtained from the suggested technique. Additionally, the suggested technique was shown appropriate for predicting the only heart disease not for other disease.

RESEARCH METHODOLOGY

This research work we focused on prediction of the three main disease which are Heart disease, Diabetes and Hypertension using the Hybrid Ensemble Common Model (HECM). That model based on combining KNN, LGBM(Light Gradient Boost) and Random Forest and Estimated with Voting classification for final result. Voting classifier worked as Meta classifier and KNN, LGBM and RF classifier will use as the base classifier and Voting Classifier used as meta classifier. dataset collected from Kaggle, different dataset taken for heart,diabetes and hypertension. All dataset have different features and impossible to join together. That why we created common model to predict with high accuracy and efficiently with less time taken. The performance of the proposed approach will be analyzed in terms of certain parameters like accuracy , precision and recall and F-measure, cross validation.

Proposed methodology Algorithm -

Following are the various steps of research methodology:-

Step 1: Input the dataset for the any disease for prediction of any disease. In our case we used heart,diabetes and hypertension dataset.

Step 2:Preprocess dataset , remove outliers, change categorical attribute into continuous variables .Split dataset 80:20 ratio, where 80 is Training dataset and 20 is test data set .

Step 3: Append classifiers on Estimator with KNN, LGBM and Random Forest combined for the prediction of the dataset, in our case Random Forest Estimator was 100 in variable. KNN has leaf_size=4, metric='minkowski',n, n_neighbors=3. And output was collected in Estimator.

Step 4: Now predict final result with majority Voting classifier with Estimator. It predict final



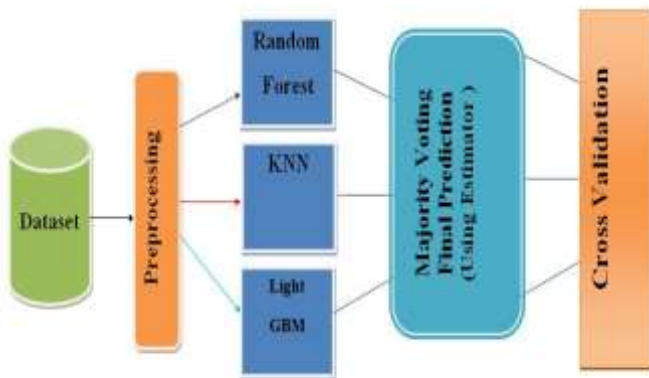
result on test dataset and return variable as predicted values.

Step 5: Cross validation done after final result at last for validate our model accuracy.

Step6: Confusion metric,Sensitivity,F1,Recall and Time taken for final prediction recorded as per our equirement.

In classification, an input is consisting of the K-nearest training examples in feature space and on the other hands, the output depends upon n classification category. The output is belongs to classification category when,

A data is classified by majority of the votes from it's neighbor with the objects being assigned to



that class which is most common it's K.N.N .

If the k=1, then the object is belongs to the class of that single nearest neighbor. KNN is the simplest algorithm in which all cases are stores and classifies based upon the similarity measures and has been used in statistical estimation and pattern recognition. Once the value for k is selected then the prediction can be made for the regression, KNN prediction is the average of KNN outcomes

$$y = 1/k \sum_{i=1}^k y_i$$

The y_i is 1st case and the y is the prediction or outcome of the query. The non-parametric approach that is used to perform classification as well as regression is known as K Near Neighbor classifier. The data is supposed to be refer in the feature space through KNN. Else within multidimensional or the scalars vectors, the given data in the applications. When the points are available in the feature space, here is the most important thing is a notion of distance. Although if we want to calculate the distance Euclidean distance is the most important algorithm, there is no need to use only this method to do calculation. When we want to estimate the density at point a, place a hypercube which centered at a and it's continuously increased till k neighbors are not captured. Then, we can apply the following formula to estimate the density:

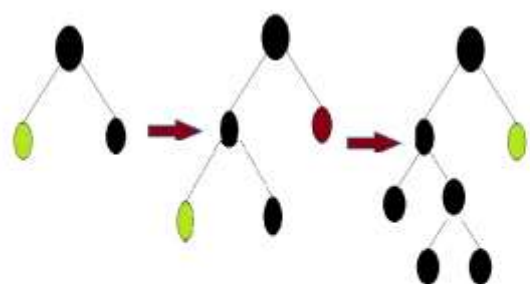
$$p(a) = \frac{k/n}{V}$$

Here, n- denotes total numbers of available data points. And V shows the volume of hypercube.

LightGBM uses the algorithm which is based on tree based learning algorithm, it's a gradient boosting framework. It's design is suitable to be distributed and efficient with the faster speed of training, low usage of memory, Better accuracy, higher efficiency. It also supports parallel and GPU learning. Light GBM increases tree vertically while other algorithm increases trees horizontally which means it (Light GBM) increases tree **leaf-wise** while other, algorithm increases level-wise. It will be able to select the leaf which has max delta loss to increase. When increasing the identical leaf, Leaf-wise algorithm is able to reduce more loss in place of level-wise algorithm.

There is given the explanation of the implementation of GBM and other boosting algorithms.





Leaf wise Tree Growth

The size of data is growing day by day and it is becoming difficult for old or ancient data science algorithms to perform faster results. Light GBM uses prefix 'Light' which denotes its high speed. It is capable to work with the large size of data. And it uses lower memory to run. Other reason of Light GBM popularity is because it is more focused on accuracy of results. Light GBM also supports GPU learning and that why scientists are widely using Light GBM for application development for data science

Random Forest, (RF) is supervised Machine Learning algorithm and it is like bagging algorithm. In early Random Forest (RF) uses Classification And Regression Tree(CART) decision tree as weak learner which was based on the GINI coefficient to select features when we generate tree. The selected features are randomly selected, so we get random result and due to randomness it is useful to variance of model. RF do not need additional pruning and have anti over fitting. There are some advantages :

- Due to ensemble classifier its accuracy is better than other single classifiers.
- It is combined tree classifier and due to this property it can classify non linear data easily.
- In training phase it learn faster than other classifier.
- it took take less time than other.

We basically need to know the impurities in dataset, so we can use Gini index which can help us to find out the impurities, we can take lowest Gini index for better results.

$$\text{Gini Index} = 1 - \sum_{l=1}^n (P_l) \times (P_l)$$

$$= 1 - [(p^+)2 + (P^-)2]$$

We can also use to measure impurity from split dataset with help of Entropy,

$$\text{Entropy}(s) = -P(+)\log_p(+)-P(-)\log_p(-)$$

Tool Description in this proposed work

1. Confusion Matrix:- The classification in matrixes learning we need to worry about the percentage of correct classification and misclassification. Then we need a method that provide accuracy and also help us to calculate correct and incorrect classifications. Here the confusion matrix come for this purpose it is generally 2x2 (NxN) matrix which solve the evaluation problem help to increase performance of matrix learning models.

Table 1: Confusion Matrix

- TP - True Positive
- TN - True Negative
- FP - False Positive
- FN - False Negative

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \dots(A.1)$$

2 Recall: The ratio of number of times the model predicts positive cases correctly to the total number of actual positive cases is known as recall.

$$\text{Recall} = \frac{TP}{TP + FN} \dots(A.2)$$

b. It is also called the F Score or the F Measure. In another word, the F1 score conveys the balance between the precision and the recall

$$F1 = 2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$$

3 . **Precision:** The ratio of number of times the model correctly predicts positive cases to the total number of positive cases predicted by it is called precision.

$$\text{Precision} = TP / (TP + FP)$$



Sensitivity:- The Sensitivity is a measurement tool which is also known as TPR (True Positive Rate) or recall.

If model have high sensitivity then the model have few false negatives in other word sensitivity is inversely proportional to false negative.

$$\text{Sensitivity} \propto 1/\text{False Negative} \dots (\text{B.1})$$

If the sum of sensitivity (TPR) and FNR would be = 1

$$\text{TPR} + \text{FNR} = 1 \dots (\text{B.2})$$

Mathematically sensitivity calculated

$$\text{Sensitivity} = (\text{TP}) / (\text{TP} + \text{FN})$$

2. True Positive :- True Positive refers that a person predicted as suffering from any disease, is actually suffering from predicted disease. In other words, true positive shows actually unhealthy person.

		Actual Values			
		Positive (1)		Negative (0)	
Predicted Values	Positive (1)	True (TP)	False (FP)	Positive	
	Negative (0)	False (FN)	True (TN)	Negative	

3. False Positive :- False Position refers that a person predicted as suffering from any disease, is a healthy person, is not actually suffering from that kind of disease.

In real time prediction – it is very deadly to person to predict him healthy but really he is unhealthy. The specificity and accuracy need to improved.

RESULT AND DISCUSSION

In this work we used spyder 4 (Anaconda) which use Python interpreter. Python is an open source and available freely for everyone. Python have large machine learning libraries like pandas, numpy and sklearn. Datasets are collected from the kaggle. The HRFLM classifier with base classifier as RF is applied in the previous research

Variables	Definition
Id	Patient ID
Gender	Gender of Patient
Age	Age of patient
Hypertension	0 - no hypertension, 1 - suffering from hypertension
Ever_married	Yes/No
Work_type	Type of occupation
Residence_type	Area type of residence (Urban/Rural)
Avg_glucose_level	Average of Glucose level (measured after meal)
Bmi	Body mass index
Smoking_status	Patient's smoking status

work for the Heart Disease prediction. In this approach dataset is given KNN, LGBM and RF in form of estimator [] .ensemble classifier (HRFLM) as input, and this generate an output which give as input to voting classifier which is applied for the final output. In this proposed method HRFLM classification method which is the combination of KNN, LGBM and RF. This HRFLM method will reduces the time complexity of the model and increases accuracy and other parameters like recall, precision , sensitivity used.

Dataset:



Table 2: Heart Disease Dataset,

Attribute	Description
age	age in years
sex	sex - (1 = male; 0 = female)
cp	chest pain type
resttbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholesterol in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	slope of the peak exercise ST segment
Ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
target	have disease or not (1=yes, 0=no)

- Heart** In our approach dataset taken from kaggle, UCI repository. For heart disease the dataset we use the well known Cleveland dataset which is collected from a UCI machine learning repository. After collecting various records data is pre-processed. There is 303 patient records, in dataset where 6 records have no values. Those 6 records are already removed from the dataset and the remaining 297 patient records are used in pre-processing. The binary classification and multiclass variable are introduced for the attributes of the given dataset. The dataset description is given in Table 2.
- Diabetes**, the PIMA Indians Diabetes dataset taken from kaggle and it contains 768 female diabetic patients from the Pima Indian population. This dataset contains 268 diabetic patients which are positive and 500 patients are negative. Dataset has 8 different attributes, data description in table 3.

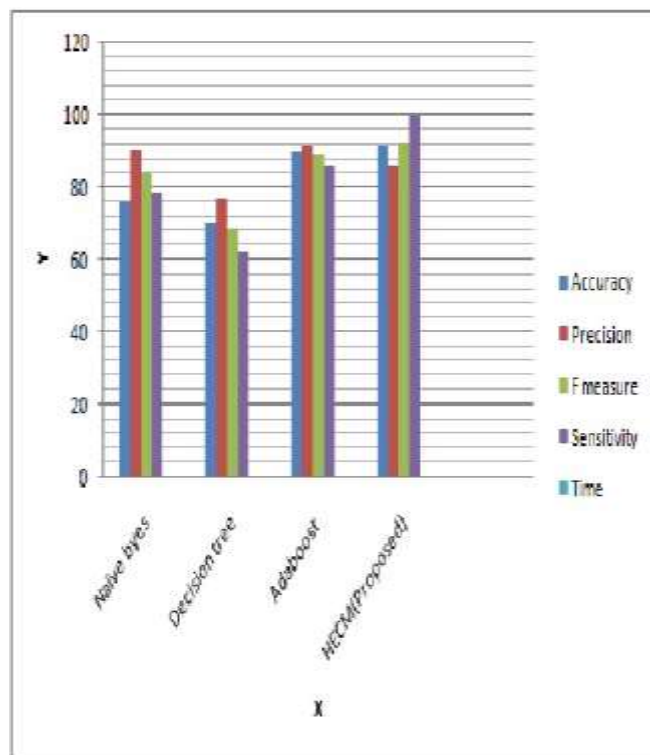
Table 3: Showing Diabetes Dataset.

- Hypertension**, The variable information related to this diseases is taken from a third-party website, the international challenge on the famous internet platform Kaggle , which gives data in the title of “healthcare data” that was uploaded by Miss. Agarwal. The data is consist with 43,400 patients and nine attributes, including the attribute which was predicted. The organizers provided the types of a data stream to help a large set of individual factors, with algorithmic development

For each dataset we followed below proposed method for classification,

Step 1: Input the dataset for the any disease for prediction of any disease. In our case we used heart,diabetes and hypertension dataset.

Step 2:Preprocess dataset , remove outliers, change categorical attribute into continuous variables .Split dataset 80:20 ratio, where 80 is Training dataset and 20 is test data set .



Step 3: Append classifiers on Estimator with KNN, LGBM and Random Forest combined for the



prediction of the dataset, in our case Random Forest Estimator was 100 in variable. KNN has leaf_size=4, metric='minkowski',n, n_neighbors=3. And output was collected in Estimator.

Step 4: Now predict final result with majority Voting classifier with Estimator. It predict final result on test dataset and return variable as predicted values.

Step 5: Cross validation done after final result at last for validate our model accuracy.

Step6: Confusion metric,Sensitivity,F1,Recall and Time taken for final prediction recorded as per our requirement.

Table 4: Heart Disease with different parameters.

Heart Disease	Model	Accuracy	Precision	F measure	Sensitivity	Time
	Naive byes	75.8	90.5	84	79	NA
	Decision tree	70	76.92	68.9	62.5	NA
	Adaboost	90	92	88.8	85.7	NA
	HECM(Proposed)	91.80	85.92	92.06	100	0.469 S

In Table 4, represent comparison of several Classifiers with our proposed HECM method . In this comparison the heart disease dataset and different machine learning algorithms are compared. Table 4 shows that the naïve byes accuracy 78.8% and Adaboost accuracy is 90% whereas our proposed HECM method showed 91.80 %. Sensitivity is higher among all classifiers. Time is also calculated and it

shown 0.47 second but as previous study no time comparison done at that time.

Fig.1. the graph plotted between accuracy and other measuring tools.

Table 5:Showing diabetes disease with different parameters

In Table 5, we also mentioned author and previous disease results. This table shows that SVM and Swarm Intelligence method which was hybrid of xgboost and naïve bayes achieved low accuracy respectively 69% and 70%. Hybrid ANN showed 77% while our proposed HECM model achieved 81.82% accuracy with 80% cross validation,70.21% precision, f1 and sensitivity. In previous work they did not calculated time taken by classifier till final classification.

In Fig.2. the graph plotted between accuracy and other measuring tools for diabetes disease.

Diabetes	Author	Model	Accuracy	Precision	F measure	Sensitivity	Time
	Baniibrata Paul	Hyb.ANN	77	76.92	76	76	NA
	Narendra Mohan	SVM	69	69.2	NA	NA	NA
	C.Kalpana	(through Swarm Intelligence)	70	77	NA	NA	NA
	Proposed 2022	HECM(Proposed)	81.82	70.21	70.21	70.21	0.61



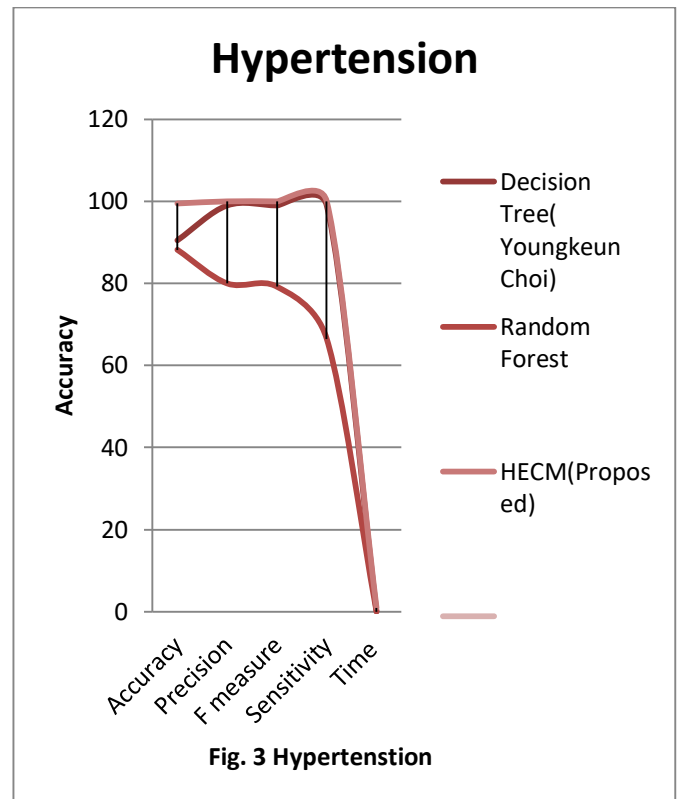
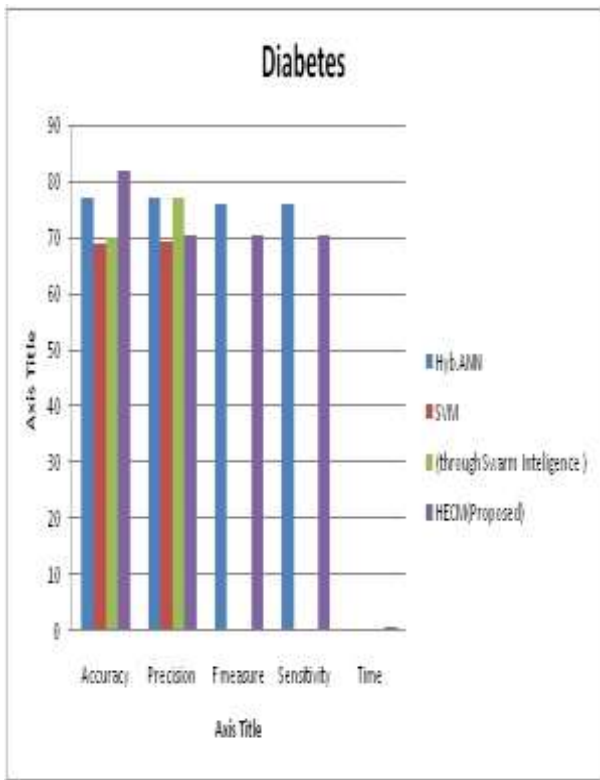


Fig. 3 Hypertension

Fig.3. The graph plotted between accuracy and other measuring tools for diabetes disease. 1599

	Model	Accuracy	Precision	F-measure	Sensitivity	Time
Hypertension	Decision Tree (Youngkeun Choi)	90.50	99	99	99	NA
	Random Forest	88.2	80	79.2	66.5	NA
	HECM(Proposed)	99.48	100	100	100	0.94

Table 6 :Hypertension Disease comparison with different parameters.

In this table we compared mainly 2 classifier with our model. Decision Tree done by youngkeun choi [22] achieved 90.50% accuracy with 99 % other parameters. Random forest shows 88.2% accuracy, F1 was 79% and sensitivity was just 66.5 % while our proposed model HECM showed 99.48% accuracy with 99% cross validation and 100% with other parameters. Hypertension prediction got the best result with this model. Time was calculated and just 0.94 second taken for final prediction.

Conclusion

In this work, we introduced a Hybrid Ensemble Common Model (HECM) for was based on supervised classifiers e.g. Random Forest, Light GBM and KNN also as Estimator or final classification done by majority Voting. In previous research work they used either single or combination of two classifiers. Also they didn't recorded time taken for classification done by classifiers. The average accuracy was 80% in previous work which was good but still need higher accuracy and need single model for classification of almost all diseases. Time is key factor while working with real time frame work or working with web based or cloud based classification, due to slow classification web server can crash or lead to delay in classification. Our proposed HECM model took 0.49 to 0.9 second for final prediction when working with dataset have tuples 1K to 10K.

The HECM achieved 99.48% accuracy with Hypertension dataset, and 100% with F1,Sensitivity and Precision. While heart disease dataset, HECM showed 91.80% accuracy.



Diabetes disease prediction we got 81.82% accuracy which was good. All disease achieved higher accuracy than previous work.

The performance of proposed model HECM is tested in terms of accuracy and other parameters, HECM model showed higher accuracy as compared to existing techniques.

In future we will add more diseases and will increase more accuracy with less time.

References

- [1] D. Ratnam, P. HimaBindu, V. MallikSai, S. P. Rama Devi and P. Raghavendra Rao, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm", 2014, International Journal of Computer Science and Information Technologies, vol. 5, no. 2, pp.2384- 2388
- [2] T. Santhanam and E. P. Ephzibah, "Heart Disease Prediction Using Hybrid Genetic Fuzzy Model", 2015, Indian Journal of Science and Technology, vol.8, no. 9, pp.797–803
- [3] G. Purusothaman and P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", 2015, Indian Journal of Science and Technology, vol. 8, no. 12
- [4] K. Srinivas, G. RaghavendraRao and A. Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", 2010, Proceedings of 5th International Conference on Computer Science & Education, China, pp. 24–27
- [5] K. Thenmozhi and P. Deepika, "Heart Disease Prediction Using Classification with Different Decision Tree Techniques", 2014, International Journal of Engineering Research and General Science, vol. 2, no. 6, pp.6-11
- [6] S. Chaitrali, Dangare and S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", 2012, International Journal of Computer Applications, vol. 47, no. 10, pp. 0975 –888
- [7] J Peter and K. Somasundaram, "An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques", 2012, Proceedings of IEEE International Conference on Advances in Engineering, Science and Management (ICAESM), pp. 514-518
- [8] Senthil kumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", 2019, IEEE Access
- [9] Sanchayita Dhar, Krishna Roy, Tanusree Dey, Pritha Datta, Ankur Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases", 2018, 4th International Conference on Computing Communication and Automation (ICCCA)
- [10] Abdelmegeid Amin Ali, Hassan Shaban Hassan, Eman M. Anwar, "Heart Diseases Diagnosis based on a Novel Convolution Neural Network and Gate Recurrent Unit Technique", 2020, 12th International Conference on Electrical Engineering (ICEENG)
- [11] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, "Classification models for heart disease prediction using feature selection and PCA", 2020, Informatics in Medicine Unlocked
- [12] InduYekkala, Sunanda Dixit, M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization", 2017, International Conference On Smart Technologies For Smart Nation (SmartTechCon)
- [13] Mohini Chakarverti, Saumya Yadav, Rajiv Rajan, "Classification Technique for Heart Disease Prediction in Data Mining", 2019, 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICT)
- [14] Xu Wenxin, "Heart Disease Prediction Model Based on Model Ensemble", 2020, 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)



[15] Norma Latif, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension", 2019, Vol 7, IEEE Access. Digital Object Identifier .1109/ACCESS.2019.2945129

[24]M.J. Shivambigai, "Hypertension Risk Prediction Using Neural Network",2021, Springer Nature 2021.

[16] NIKOS FAZAKIS," Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction",2021 IEEE Access, Vol. 9. Digital Object Identifier 0.1109/ACCESS.2021.3098691

[17] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1-3, doi: 10.1109/ICECA49313.2020.9297411.

[18] B. Paul and B. Karn, "Diabetes Mellitus Prediction using Hybrid Artificial Neural Network," 2021 IEEE Bombay Section Signature Conference (IBSSC), 2021, pp. 1-5, doi: 10.1109/IBSSC53889.2021.9673397.

[19] Silva, G.F.S., Fagundes, T.P., Teixeira, B.C. et al. Machine Learning for Hypertension Prediction: a Systematic Review. *Curr Hypertens Rep* (2022).

[20] A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," in *IEEE Access*, vol. 9, pp. 39707-39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[21] W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou and W. Cao, 2019 "A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease," in *IEEE Access*, vol. 7, pp. 175248-175258, 2019.

[22] Choi, Youngkeun and Jae Choi. "Hypertension Prediction Using Machine Learning Technique." *IJSDS* vol.11, no.3 2020: pp.52-62. <http://doi.org/10.4018/IJSDS.2020070103>

[23] D. LaFreniere, F. Zulkernine, D. Barber and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*.

