DISEASE PREDICTION BASED ON SYMPTOMS USING ENSEMBLE AND HYBRID MACHINE LEARNING MODELS

Bharathi Mohan G.
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
g_bharathimohan@ch.amrita.edu

Prasanna Kumar R.
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
r_prasannaKumar@ch.amrita.edu

Harsha Vardhan
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
akulaharsha1435@gmail.com

Veda Sampreetha M.
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
veda.maganti@gmail.com

Siva Sairam A.
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
adapasairam090904@gmail.com

Syfullah sd.
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
syedsyfullah4@gmail.com

Sriram K.
department of CSE
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
kdvsriram@gmail.com

Yamani Kakarla
Business Data Scientist
Halifax, Nova Scotia
Canada
yamanikakarla@gmail.com

***Abstract*—The health sector is one of the prominent sectors in which continuous advances are happening to meet present population requirements as well as to withstand new diseases. As the population increases analyzing patients' data becomes a huge task as big data analytics helps full to manage and track those data efficiently lack of infrastructure in many areas cause providing health care a challenging task due to the constant change in the environment leads to many health crises even predicting disease based on symptoms at early stages by the doctors are less accurate because continuous mutations are also happening in disease-causing agents which makes most similar symptoms and difficult to predict. As the health sector consists of a huge amount of data, prediction of disease building a Machine learning (ML) model will give more accurate results. In this research, hybrid ml model (Naïve bayes + Ada boost) and Ensemble ml model(Random forest + Gradient boosting) and the final prediction is done by Voting classifier this model is also termed as voting Ensemble, showing huge promises and potential for accurate prediction than ML models.Performed 10 fold Cross Validation and can attain an accuracy of 85% for Naive Bayes + Ada Boost and 92% for Random forest + Gradient boosting.**

***Index Terms*—Hybrid Machine learning,Ensemble learning, Random forest, Gradient Boosting, Naive Bayes, Ada Boost, Clustering, Hyper parameter Tuning, Cross Validation, Boot Strapping, Blending**

I. Introduction

The stable development of innovation in the area of healthcare is necessary to meet the growing demands of an increasing global population and the outbreak of diseases. The medical field is at the forefront of these problems and handling and evaluating the patient data has increased to an extent. Big data analytics [1] is now a necessary tool for handling this influx of healthcare data properly and also for allowing the accurate monitoring of medical outcomes. Providing medical services is challenging because the medical system is not homogeneous and poor infrastructure is present in many regions.

Regardless of all these obstacles medical industry takes benefits from many Machine Learning algorithms. ML provides us with an increasingly accurate prediction of the disease by analyzing and processing a large amount of data. This paper provides us with the potential of Machine Learning in predicting the accuracy of illness.Machine learning is a field of AI in which algorithms are build to perform specific tasks which are trained by a proper data set. The designed model can learn from the training and executes for test data based on algorithms [2] on which the model has built. One of the main Key component of machine learning is Data it is a core of any model that has been built. Quantity as well as Quality of the dataset decides the performance of the model. Many such models are build and introduced and the advancements are still going on like Auto ML [3]. Decision Tree ,Random forest ,Support Vector machine(SVM) ,K-Nearest Neighbour(KNN) these are few algorithms that are mostly build to predict diseases based on symptoms.Making a model to learn some thing can be supervised or unsupervised. In supervised learning algorithms like KNN ,Random forest ,Decision Tree ,SVM etc... labels are provided to the data and feeds to the algorithm which maps inputs to outputs which makes to predict unknown data. Where as in unsupervised learning data is feeded the algorithm should itself need to extract the features and finds the patters from the data with out any labelling using some techniques like clustering [4], Word embedding, LDA(Latent

Dirichlet Allocation) etc… in which data is grouped based on similar characteristics and patters. Using KNN and convolutional neural networks(CNN) models are proposed based on word embedding and common approach for classification. ID3 algorithm followed decision tree, Random forest, SVM with Radial Basic Function as Kernel function and XGboost with decision tree and the models which are build and proposed from past 15 years even though lack of proper dataset, limited evaluation and performance metrics few models are not generalised ,models that are proposed from 2012 have better datasets and performance metrics even lack of validation techniques models may undergo over fitting. And results getting from sing ML algorithm is not upto the mark for few applications optimizer are introduced to increase the performance of ML algorithms but models cannot able to deals with application where it requires both structured and unstructured data to build and execute [5]. Here comes the origin of Hybrid models ,which Combines multiple machine learning algorithms with best parameters makes hybrid machine learning models [6]. These improves the strength ,efficiency ,performance of prediction ,robustness. The main aim of introducing hybrid models is not only to increase efficiency and robustness these models can able to solve many problem statements which cannot be done by optimizer and other algorithms.

This project consists of the development of hybrid machine learning models that connect various ML model, such as the fusion of Naïve bayes with Ada boost .These hybrid models are the cutting-edge approach that improves the prediction of illnesses and overcomes one model limitation by another algorithm which improves efficiency and reduce fault tolerance.And Voting Ensemble model i.e combining Random forest and Gradient Boosting with voting Classifier,This combination is also known as Blending we calculated Accuracy, F1 score and precision as performance metrics for both training and validation by taking average of 5 iteration of the model to reduce fault tolerance and make the results more generalise and performed 10 fold cross validation.

II. LITERATURE SURVEY

"Human symptoms disease network" [7] an Article which was published on 26th jan 2014 written by Xue Zhong Zhou ,Jorg Menche ,Albert-Laszlo Barabasi and Amitab sharma.They constructed a network which was based on disease symptoms to investigate the connections and molecular interactions which are underlying related protein interaction patterns [8] by clinical manifestations of disease.Constructed Human symptoms disease network(HSDN) with 7,109,429 records from pubmed which yields 4,422 diseases and 322 symptoms. They statistically analysed symptoms and diseases which shared symptoms indicate shared gene as well as shared protein interactions.Evaluated symptom similarities, Graphically analysed disease similarity bins with Human – phenotype ontology(HPO) network [9] overlapping. And found correlation between symptom similarities and shared gene. HSDN model is analysed by shortest path and single linkage between disease modules ,Disease diversity to characterise the connections of node with in network. HSDN has 84% of pairwise disease links. They also determined statistical significance to get accurate value of similarities for pair wise symptoms. For performance evaluation they did extensive manual check.

Dhiraj Dahiwade ,prof.Ganjanan patle ,prof.Ekta Meshrom [10] designed K- Nearest neighbour(KNN) [11] and Convolutional Neural Network(CNN) [12] models to predict diseases based on symptoms in 2019 with patient disease dataset from University of California Irvine Machine learning repository. Compared and concluded CNN is better with accuracy of 84.5% ,The overall CNN model is just a common approach for classification of text and word embeddings there are many developments done from this basic pipeline one of those is Recurrent neural networks(RNN) [13] which is more complex and better. Other than accuracy they did not go with other metrics like fault tolerance , error and there are many limitations validation and generalization of model performance is tough with out cross validation [14].

Sneha Grampurohit ,Chetan sagarnal [15] designed a prediction model using existing ML algorithms Decision tree ,Random Forest ,Naïve Bayes. In their work they built a GUI that displays the predicted disease based on user given symptoms by decision tree, Random forest and Naive bayes. They considered data set of 132 symptoms and 41 diseases with 4920 records. For decision tree they used ID3 algorithm [16] which follows top down greedy search through given columns. Mathematically described how algorithms work as a result they obtained accuracy as 0.95 for all three algorithms. Limitations of this model is that with out any optimiser Decision tree is a week algorithm to classify and they did not generalised the model. They used very small data which may leads to over fitting as well as with present huge amount of medical data this model fails to predict.

"Multiple disease prediction using ML algorithms" [17] a paper published in 2021 by Orlando Leiva – chauca, Antonio Huaman – Osorio ,Tatiana Gonzaled – Yanac. They build ML model with Decision tree ,Support Vector Machine(SVM) and Naïve bayes used Radial Basic Function(RBF) as Kernal function [18] for SVM there are many such Polynomial Kernel function(PKF) ,Exponential Radial Basis Kernel function(ERBKF) ,Linear Kernel function(LKF). RBF is identified as best because of its unique characteristics like Implicit feature mapping ,Universal Approximation [19] ,smooth Decision boundaries. They attain accuracy of 77% for naïve Bayes ,86% for SVM and 90% for Decision tree ,But algorithm ID3 which they used for decision tree is common there are many advanced algorithms like c4.5 and c5.0 which are successor to ID3 and in decision tree there is a chance of over fitting and with ID3 it is week classifier compared with others.

Carlos Andre's Tavera Romero along with his team proposed an Article "Early – stage Alzheimer's Disease prediction using Machine learning " [20] on march 2022 .They proposed ML model using Decision tree, Random forest and SVM. They used extreme gradient Boost along with decision tree. As the Decision tree is a week classifier by adding XGBoost improves its strength to predict and it increases speed and performance.After prediction they go for voting [21] [22] which is not actually a classifier but bring it is to harness the collective intelligence of different models which improves accuracy, prediction. For model validation they performed 10 fold cross validation and calculated accuracy ,precision ,Recall ,F1 score as performance metrics, Results of their work as follow 80% accuracy in Decision tree ,97% for Random forest ,81% for SVM and 95% for XGBoost. Random Forest is better algorithm for predicting Alzheimer's Disease. Dataset used from Kaggle having cross – selection Magnetic Resonance imaging(MRI) data in young, middle ,aged ,Non – demented and demented older adults and Longitudinal MRI data in Non – demented and demented older adults.

Senthil kumar mohan ,Gautam Srivastav [23] proposed an Hybrid machine learning model to predict Heart Disease [24] in 2019. They proposed Hybrid random forest with linear model(HRFLM) with accuracy 88.7%. They also done comparative analysis with different ML models. ANN with back propagation [25] is done in HRFLM with 13 features as input previously an hybrid model Recurrent Fuzzy neural network(RFNN) [[26] is introduced with a combination of genetic algorithm and fuzzy NN for Heart disease diagnosis. HRFLM has 4 algorithms embedded in it Algorithm 1: Partition based on Decision tree , Algorithm 2: For less error rate apply Machine learning Algorithm 3: Using less error classifier for extracting Features Algorithm 4: Apply classifier on extracted Features.Used Rattle which is a GUI (Graphical user interface) package in R for visually analyse the data

III. OVERVIEW AND PROPOSED MODELS

Dataset is collection of disease symptoms based on the information at New York Presbyterian Hospital. Contains 150 different diseases and 132 frequent occurring Symptoms.

Dataset is trained and evaluated with 3 selected algorithms in which 2 are hybrid algorithms
a) Random Forest
b) Naïve Bayes + Ada Boost
c) Random Forest + Gradient Boosting

A. Random Forest

Random forest is one of a better machine learning algorithm even with minimal tuning of Hyper parameters. It is an ensemble of decision tree with unique characteristics to migrate the problem of over fitting. In general it can also comes under hybrid machine learning algorithm as it is a combination of Decision tree and voting classifiers.

Being more robust even as per previous proposed algorithms Random forest has better accuracy and consistency, Making it as the base model and understand the architecture shown in **Figure.1** and analyses the drawbacks of it and making an advanced ensemble model to overcome the drawback which improves prediction and reduce bias and variance.

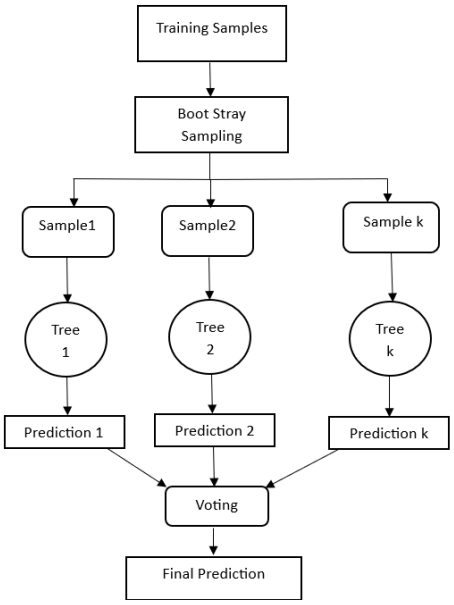

Fig. 1. Random Forest Architecture

Architecture: Performs Boot Strapping to select set of data randomly and generates multiple samples which are used by different decision tress to train. For splitting data a subset of features are to be selected at each node of Decision tree. To select those random forest introduce randomness. On the basis of selected features for each sample a decision tree is constructed. Minimum samples per leaf, maximum depth these hyper parameters will define stopping criterion for growing tree. each decision tree predict separately. After prediction every decision tree votes for a class ,final prediction is the one which has highest number of votes. To overcome over fitting and to make a generalized model random forest follows ensemble approach. Even though it follows ensemble approach when it comes foe complex data the memory usage is limited and random forest may leads to be bias and variance.

B. Naïve Bayes + Ada Boost

Naïve Bayes is based on Bayes'theorem which is a probabilistic algorithm. Ada boost is an ensemble method, Combines weak classifiers to generate a strong classifier. Combing these two will be a hybrid model which increase the entire efficiency and performance.

Work flow: To begin, train many Naive Bayes classifiers. A probabilistic classifier called Naive Bayes determines the conditional probability of a class given some characteristics. It makes the simplifying (naive) assumption that features are conditionally independent, given the class, which frequently works well in practise.An individual subset of the training data is used to train each Naive Bayes classifier. These subsets can be produced using alternative techniques or random sampling with replacement (bootstrap samples). The performance of each Naive Bayes classifier is evaluated using the training data. We calculate the accuracy or error rate, for each classifier. Classifiers that perform better are assigned weights while classifiers, with performance receive lower weights. The specific weight distribution is determined based on the error rate of each classifier. Creating a Weighted Ensemble (Ada Boost)The trained Naïve Bayes classifier is get combined with Ada Boost algorithm into a weighted ensemble. As the weights in the ensemble is higher the performance of classifier is better.

C. Random forest + Gradient boosting

Figure. 2 represents the workflow of Random forest + Gradient boosting algorithm an Ensemble model. And the detailed explanation is described in coming paragraph.

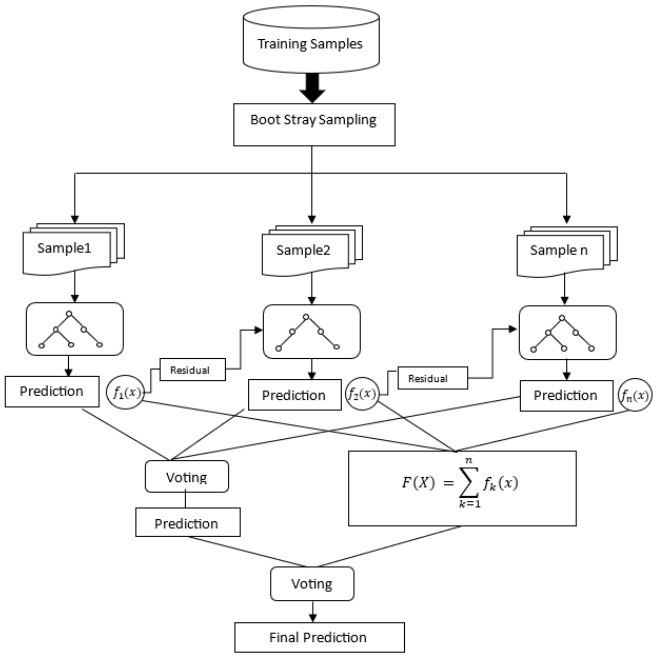

Fig. 2. Model Architecture

Random forest is used for both classification and regression and this model is known for its robustness and ability to handle large amount of data this method works for both numerical and categorical data. Gradient boosting is used for more accuracy over random forest as it can handle more complex data as it is a model which tries to correct the errors of the previous ones and this model focuses on minimizing the prediction errors. Even though those two algorithms are prone to over fitting, Using Random Forest and Gradient Boosting model combined is a type of model stacking which reduce over fitting increase model interpretability and Flexible Hyper parameter Tuning and even can deal Complex data. To build the combined ensemble model the first step is data preparation i.e gather the data, pre-process the data and split it into training and testing dataset. Training data is divided into n number of samples using Boot strapping for each bootstrap sample a decision tree has been built. For every decision tree there will be a prediction and all those n predictions from n decision trees are feeded to voting classifier to give a final prediction termed as Prediction, As it comes to Gradient boosting we initialize the model, then we calculate the residuals for each decision tree and those samples are passed to next tree to rectify the errors and increase the prediction accuracy of the base model that is used and the final output we get F(x). Final output will be decided by voting classifier. The combination of these algorithms increase prediction accuracy, reduce over fitting and even benefits as random forest provides robustness and gradient boost focuses on rectifying the errors done by the base model and reduce the bias. And even can handle huge amount of data.

IV. Methodology

As given in the Figure.3 we implemented the workflow to build the model. The data with 150 diseases and 132 symptoms is collected and pre-processed i.e dealing with missing values and getting ratio of null values. Defines target variable y – 'prognosis' and input variables x – 'Symptoms' Splitting the data 80% for training and 20% for testing. Deploy model with hybrid algorithm (Navie Bayes + AdaBoost) and a model with combination of voting Ensemble algorithm (Random forest + Gradient boost) and train with the data which is already split. For any machine learning models there should be some parameters in which it performance depends on termed as Hyper parameters.

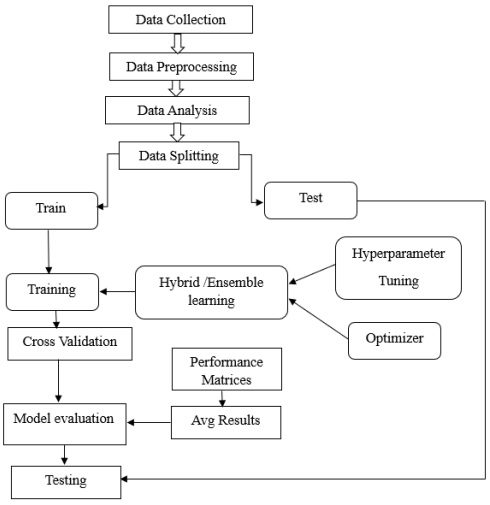

Fig. 3. Model Workflow

To select best hyper parameters ,performed a Hyper parameter tuning method “GridSearchCV”.It can choose the best set of parameters by evaluating model performance with all combinations from parameter grid. The best hyper parameters of algorithms are provided in Table I

TABLE I. Best Hyper parameters

Algorithm	Hyper parameters	Values
Random Forest	Min_samples_leaf N_estimators	10 200
Gradient Boost	Learning Rate Max_depth N_estimators	0.1 6 300
Naïve Bayes with AdaBoost	N_estimator	300

V. Model Evaluation and Result Analysis

After getting best Hyper parameters of particular algorithm model is trained and evaluated. For training and validation data accuracy, F1 Score and precision are calculated for 5 runs and average is considered for Training and validation data sets which reduces over fitting or under fitting, improves generalization and make the model more robust. Table II. shows the values of performance metrics for each algorithm. As per the Figure.4 about the confusion matrix for Random

TABLE II. PERFORMANCE METRICS

Model	F1 Score	Precision	Accuracy
Random Forest	0.88	0.88	0.89
NaïveBayes+ AdaBoost	0.85	0.84	0.85
Random Forest + Gradient Boost	0.91	0.92	0.92

Forest the model predict 176 correct out of 200 records which is 88% and 24 in correct predictions

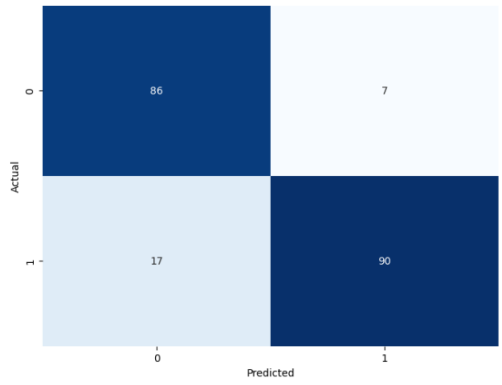

Fig. 4. Confusion matrix for Random Forest

In Fig.4 of confusion matrix for Random Forest out of 200, 176 records are predicted truly by Random forest where as 24 records are predicted false.

In Figure.5 we observe 159 out of 200 records are predicted truly by Naive Bayes + Ada Boost Hybrid model and 41 records are predicted false.

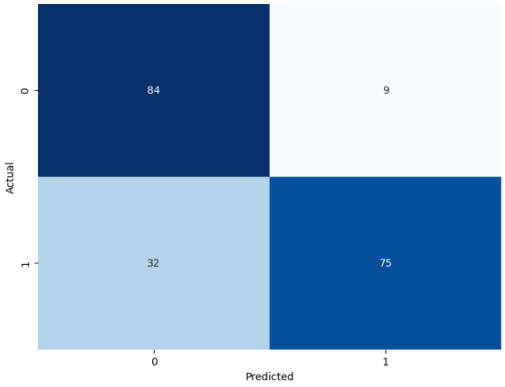

Fig. 5. Confusion matrix for Naive Bayes+AdaBoost

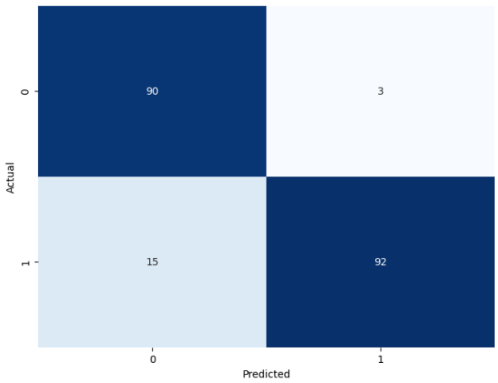

Fig. 6. Confusion matrix for Random Forest+Gradient boost

Similarly in Fig.6 182 out of 200 records are predicted correct by Random Forest + Gradient boost and 18 are incorrect prediction.

If we compare all the confusion matrices that is of Random forest, Naive Bayes + Ada Boost Hybrid Model and Random Forest + Gradient boost ensemble model it is observed that Random Forest + Gradient boost has predicted more correctly when compared to other models. The dataset we have used is linear, for which ensemble models are more efficient when compared to hybrid models.

Table III displays the predicted outputs for the set of input features which are exactly same as true values. This observation highlights models exceptional performance and illustrates how well it fits the given dataset. The model’s accuracy and reliability are further supported by the constancy with which results are predicted, which attests to the model’s

effectiveness in capturing underlying patterns in the data.

TABLE III. Result

Input (Symptoms)	Output (Predicted)	Actual
Fatigue, nausea, Yellowish skin, Loss of appetite, Yellowing of eyes	Hepatitis C	Hepatitis C
Painful walking, Muscle weakness, Stiff neck, Swelling joints, Movement stiffness	Arthritis	Arthritis
itching, Skin rashes, nodal skin eruption, Dischromic patches	Fungal Infection	Allergy
Chills, Vomiting , High fever, Fatigue, Head ache, Nausea, Constipation ,abdominal pain	Typhoid	Typhoid

VI. CONCLUSION

Machine learning is implemented in every field to make things faster and to have higher accuracy and performance including medical field. New models are emerging for better performance and many evolution are happening in machine learning one of those notable evolution is building Hybrid machine learning models and Ensemble models, These combines multiple machine learning algorithms to overcome one model limitation by adding another model and also increases the overall performance and accuracy. This hybrid model and Ensemble model approach make a better way for the future advancement and can build more sophisticated and precise prediction models. For Symptoms based disease prediction we build a hybrid models combining navies bayes combined with AdaBoost which gives 85% accuracy and an Ensemble model combining Bagging and boosting models i.e random forest and gradient boost with voting classifier which attains accuracy of 92% ,If you go with only random forest it attains 89% accuracy. And We conclude Random forest with gradient boost is one of the best model to predict disease based on symptoms.

REFERENCES

[1] Spandana, Chereddy, Ippatapu Venkata Srisurya, S. Aasha Nandhini, R. Prasanna Kumar, G. Bharathi Mohan, and Parathasarathy Srinivasan. "An Efficient Genetic Algorithm based Auto ML Approach for Classification and Regression." In 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 371-376. IEEE, 2023.

[2] [1] Mahesh, Batta. "Machine learning algorithms-a review." International Journal of Science and Research (IJSR).[Internet] 9, no. 1 (2020): 381-386

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] [2] Mohan, G. Bharathi, R. Prasanna Kumar, and T. Ravi. "Coalescing Clustering and Classification." (2012): 446-450..

[5] Saravi, Babak, Frank Hassel, Sara Ülkümen, Alisia Zink, Veronika Shavlokhova, Sebastien Couillard-Despres, Martin Boeker, Peter Obid, and Gernot Michael Lang. "Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models." Journal of Personalized Medicine 12, no. 4 (2022): 509.

[6] Khayyam, Hamid, Ali Jamali, Alireza Bab-Hadiashar, Thomas Esch, Seeram Ramakrishna, Mahdi Jalili, and Minoo Naebe. "A novel hybrid machine learning algorithm for limited and big data modeling with application in industry 4.0." IEEE Access 8 (2020): 111381-111393.

[7] Menche, Jörg, Albert-László Barabási, XueZhong Zhou, and Amitabh Sharma. "Human symptoms–disease network." (2014).

[8] Ghiassian, Susan Dina, Jörg Menche, and Albert-László Barabási. "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome." PLoS computational biology 11, no. 4 (2015): e1004120.

[9] Crowson, Matthew G., Jonathan Ranisau, Antoine Eskander, Aaron Babier, Bin Xu, Russel R. Kahmke, Joseph M. Chen, and Timothy CY Chan. "A contemporary review of machine learning in otolaryngology–head and neck surgery." The Laryngoscope 130, no. 1 (2020): 45-51

[10] Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, 2019.

[11] Hota, Soudamini, and Sudhir Pathak. "KNN classifier based approach for multi-class sentiment analysis of twitter data." Int. J. Eng. Technol 7, no. 3 (2018): 1372-1375.

[12] Albashish, Dheeb, Rizik Al-Sayyed, Azizi Abdullah, Mohammad Hashem Ryalat, and Nedaa Ahmad Almansour. "Deep CNN model based on VGG16 for breast cancer classification." In 2021 International conference on information technology (ICIT), pp. 805-810. IEEE, 2021.

[13] Keren, Gil, and Björn Schuller. "Convolutional RNN: an enhanced model for extracting features from sequential data." In 2016 International Joint Conference on Neural Networks (IJCNN), pp. 3412-3419. IEEE, 2016.

[14] Tougui, Ilias, Abdelilah Jilbab, and Jamal El Mhamdi. "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications." Healthcare informatics research 27, no. 3 (2021): 189-199.

[15] Grampurohit, Sneha, and Chetan Sagarnal. "Disease prediction using machine learning algorithms." In 2020 International Conference for Emerging Technology (INCET), pp. 1-7. IEEE, 2020.

[16] Yang, Shuo, Jing-Zhi Guo, and Jun-Wei Jin. "An improved Id3 algorithm for medical data classification." Computers and Electrical Engineering 65 (2018): 474-487.

[17] Arumugam, K., Mohd Naved, Priyanka P. Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, and Tatiana Gonzales-Yanac. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings 80 (2023): 3682-3685.

[18] Patle, Arti, and Deepak Singh Chouhan. "SVM kernel functions for classification." In 2013 International conference on advances in technology and engineering (ICATE), pp. 1-9. IEEE, 2013.

[19] Ding, Xiaojian, Jian Liu, Fan Yang, and Jie Cao. "Random radial basis function kernel-based support vector machine." Journal of the Franklin Institute 358, no. 18 (2021): 10121-10140.

[20] Kavitha, C., Vinodhini Mani, S. R. Srividhya, Osamah Ibrahim Khalaf, and Carlos Andrés Tavera Romero. "Early-stage Alzheimer's disease prediction using machine learning models." Frontiers in public health 10 (2022): 853294.

[21] Burka, Dávid, Clemens Puppe, László Szepesváry, and Attila Tasnádi. "Voting: a machine learning approach." European Journal of Operational Research 299, no. 3 (2022): 1003-1017.

[22] Mushtaq, Zaigham, Muhammad Farhan Ramzan, Sikandar Ali, Samad Baseer, Ali Samad, and Mujtaba Husnain. "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques." Mobile Information Systems 2022 (2022): 1-16.

[23] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554.

[24] Assegie, Tsehay Admassu, Prasanna Kumar Rangarajan, Napa Komal Kumar, and Dhamodaran Vigneswari. "An empirical study on machine learning algorithms for heart disease prediction." IAES International Journal of Artificial Intelligence 11, no. 3 (2022): 1066.

[25] Siregar, Sandy Putra, and Anjar Wanto. "Analysis of artificial neural network accuracy using backpropagation algorithm in predicting process (forecasting)." IJISTECH (International Journal of Information System and Technology) 1, no. 1 (2017): 34-42.

[26] Kaur, Jagmohan, and Baljit S. Khehra. "Fuzzy logic and hybrid based approaches for the risk of heart disease detection: state-of-the-art review." Journal of The Institution of Engineers (India): Series B 103, no. 2 (2022): 681-697.