**LITERATURE REVIEW**

[1] V. Jackins1 · S. Vimal1 · M. Kaliappan2 · Mi Young Lee3,”AI‑based smart prediction of clinical disease using random forest classifier and Naive Bayes”, Department of IT, National Engineering College, Kovilpatti, Tamil Nadu, India, Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam, Tamilnadu, India, Department of Software, Sejong University, Seoul, South Korea

In the paper proposed, random forest classification algorithm is used to classify many disease datasets like diabetes, heart disease and cancer to check whether the patient is affected by that disease or not.

To check for true and false case in the training and testing data Naive Bayes algorithm is used. Also, Random forest algorithm is used on the data sets. Both of these algorithms calculates results accuracy and it gives three classes separately using confusion matrix. When these two models are compared model trained with Random forest gives accurate classification. It gives the accuracy of 74.03, 83.85 and 92.40 for diabetes, heart disease and cancer diseases respectively.

Limitation**:** Random forest algorithm is slightly influenced by noise due to classification accuracies slightly decreased with increasing levels of noise.

Overcome: Noise in the dataset can be handled by outlier treatment.

[2] Subrata Kumar Mandal, “Performance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree”, Information Technology Department, Jalpaiguri Government Engineering College, Jalpaiguri, West Bengal, India

In the proposed paper, Naïve Bayesian model is used to predict a disease based on the symptoms accurately, it is easy to build, also very helpful while handling the large dataset. It is popularly called as sophisticated classification method.

Initially disease dataset is taken as input for the system, data pre-processing is done, so that it reduces the unwanted information in the further analysis .Checking the missing values and the correlation helps in splitting the data into 70% training data, 15% testing data and 15% validation data. To check the number of true and false cases in each of the three class data, we train our three classes with Naïve Bayes algorithm and calculates the results accuracy with the three classes separately using confusion matrix.

The testing diagnosing accuracy that was the main performance measure of the classifier was about 74.24%, in comparing with other well-known machine learning models.

Limitations: The algorithm faces the ‘zero frequency problem’ where it assigns zero probability to a categorical variables whose category in the dataset was not available in the training dataset.

Overcome: This could be solved by training the model with large training data so that we see all the possible categories are included in the training dataset.

[3] Chen Wang, Man Xu, “The Research of Doctors Recommendation Algorithm based on Clustering and Collaborative Filtering”, Department of management science and Engineering, Nankai business school, Tianjin, China.

To reduce the complexity and improve operational efficiency this paper proposed an improved algorithm based on clustering and clustered doctors who are professional on certain illness. The result shows that the algorithm can recommend doctors to patients effectively. First clustering the doctors according to the illness that they are professional on, then we get the basic score using collaborative filtering methods. Collaborative filtering is used to match the user’s interests.

Limitations: The recommendation in original doctor rating system may be inaccurate and unreasonable and also rating a new doctor.

The proposed system can best overcome these shortcomings.

Overcome: The recommendation system provides two types of doctors as output.

1. Doctors with respect to patients’ rating as well as their nearby location
2. Doctors with respect to users nearby location, so that new doctor can be considered based on location.

Since the patients’ score for doctors collected by hospitals are limited, we try to use the collaborative filtering to predict the score for doctors.

[4] Kriti Gandhi1, Mansi Mittal2, Neha Gupta3, Shafali Dhall4 1, 2, 3, 4,” Disease Prediction using Machine Learning“, Department of Information Technology, Bharati Vidyapeeth’s College of Engineering, New Delhi, India

In this paper, simple classified model is built using the concepts like sklearn library. Scikit-learn is one of the machine learning library used in Python. It uses various algorithms like support vector machines, random forests, and k-neighbours and it supports Python numerical and a concept called scientific libraries which include NumPy and SciPy. The model is compiled and the next step is to train it and then fit the model on the respective data using the fit() function and the model is evaluated.

The model is evaluated using various machine learning algorithms like KNN (K-Nearest Neighbours), Decision tree, Naïve Bayes, Random Forest on the dataset. The dataset used comprises of 133 columns, comprising of 132 varied symptoms experienced by patients. A total of 40 diseases are present in this dataset. The results obtained by the different algorithms used for prediction of disease revealed that Random Forest has the highest performance followed by KNN, Naive Bayes, SVM, decision tree with accuracy of 96.01, 95.99, 95.62, 95.05 and 80.85 respectively.

Limitations: The model which fits on the original dataset comprising of all the features might lead to an unwanted situation like overfitting.

Overcome: The data simplification method can be used to reduce overfitting by decreasing the complexity of the model to make it simple enough that it does not overfit.

[5] Md.Zahangir Alam, M. Saifur Rahman, M. Sohel Rahman, “A random forest based predictor for medical data classification using feature ranking” Dept of CSE, BUET, ECE Building, West Palasi, Dhaka,1205, Bangladesh.

In this paper proposed, feature ranking algorithms are used to rank features and top ranked features are selected. Then Random forest algorithm is used to train and construct final model. For each of the disease three separate subsets of highly ranked features and three separate models are constructed along with base model which includes model with all features in it. Performances are compared and best model is chosen here. Only for 10 diseases these models are constructed and checked here.

Extensive experiments are also done with other classifiers like Support machine vector, Bayes Network, Multilayer Perceptron, etc. and found the better contribution of random forest across all datasets as compared to other classifiers.

Limitation: Feature ranking do not evaluate the quality more than one feature at a time and also it does not guarantee performance increase always.

Overcome: Constructing single base model using random forest to predict the results without using feature ranking algorithm

[6] Rudra A. Godse1, Smita S. Gunjal2, Karan A. Jagtap3, Neha S. Mahamuni4, Prof. Suchita Wankhade5 “Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively”, International Journal of Advanced Research in Computer and Communication Engineering, Pune, India.

In this paper, for small problems, the users have to go personally to the hospital for check-up which is more time consuming. Also handling the telephonic calls for appointments is quite hectic. Such a problem can be solved by using disease prediction application by giving proper guidance regarding healthy living.

This paper concentrate on providing immediate and accurate disease prediction to the users about the symptoms they enter along with the severity of disease predicted.

The main feature will be the machine learning, in which algorithms such as Naïve Bayes Algorithm, K-Nearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine, which will help in getting accurate predictions and Also, will find which algorithm gives a faster and efficient result by comparatively-comparing.

Limitation: The present approach of the systems focuses only on automating the process which lacks in building the user’s trust in the system.

Overcome: By providing Doctor’s recommendation in our system, we ensure user’s trust side by side ensuring that the Doctor’s will not feel that their Business is getting affected due to this System.

[7] Ankita Tyagi, Ritika Mehra, Aditya Saxena,” Interactive Thyroid Disease Prediction System Using Machine Learning Technique”, 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, India

In this paper, different machine learning techniques are proposed for the diagnosis and prevention of thyroid disease. Machine Learning Algorithms, support vector machine (SVM), K-NN, Decision Trees were used to predict the estimated risk on a patient’s chance of obtaining thyroid disease.

Initially, a data related to thyroid of a patient is collected. Features are then extracted and selected according to the need. The data is then divided into training and testing data. Those data are then classified and predicted using different algorithms by using the knowledge base. A comparative analysis of the different algorithms is done using the performance metrices, then the prediction is done on the basis of performance.

Attributes like Age, Sex, TSH (Thyroid – Stimulating Hormone), T3 (Triiodothyronine), TBG (Thyroid binding globulin), T4U (Thyroxin utilization rate), TT4 (Total Thyroxin), FTI (Free Thyroxin Index) are considered. Algorithms like Artificial Neural Network, Support Vector Machine, Decision Tree, K- Nearest Neighbor are applied and accuracy of 97.50, 98.62, 99.63, 75.76 are observed respectively.

Limitation: It uses a lot of parameters of patients for diagnosis of thyroid disease. More parameter mean a patient has to undergo a greater number of clinical tests which is both cost effective as well time consuming.

Overcome: By considering only the observed symptoms of the patient, the number of parameters is reduced thereby the reducing the cost and time.

[8] K.M Al-Aidaroos, A. A Babar and Z. Othman, “Medical data classification with naïve bayes Approach” Centre for artificial intelligence technology, Faculty of information and science technology, Selangor, Malaysia.

In the proposed paper, NB has proven its effective application often reported as surprisingly accurate in text classification, medical diagnosis. Comparing to other classifiers NB has proven simple, efficient and also don’t require lot of parameter and naturally robust.

One if the main advantage of NB approach is that Naïve bayes classifier takes into account evidence from many attributes to make final prediction and provide transparent explanations of its decisions and therefore classified as the one of the best classifier to support medical decisions.

In the comparative study of the discretization methods for medical data mining, NB seems to be best classifier comparing to other variations of it. The bad performance of Naïve bayes is in terms of choice of dataset which is reduced by training dataset without any noise.

Limitations: In the real time systems missing values are frequently present and most of the patients records lack certain data. This may result in incorrectness, inconsistency in the model.

Overcome: This can be handled by pre-processing the dataset to reduce the missing values by various techniques like imputation, deletion, outlier treatment etc.