Vol. 3, No. 1; Jan – Mar (2024)

Quing: International Journal of Innovative Research in Science and Engineering

Available at https://quingpublications.com/journals/ijirse

Disease Prediction using Machine Learning Techniques

Shraddha Mahapatra
UG Students, School of Computer Science and Engineering, VIT University, Chennai, TN, IND.

Riddhi Bandyopadhyay
UG Students, School of Computer Science and Engineering, VIT University, Chennai, TN, IND.

Paridhi Rathore
UG Students, School of Computer Science and Engineering, VIT University, Chennai, TN, IND.

Dr. E. Elakiya★
Assistant Professor, School of Computer Science and Engineering, VIT University, Chennai, TN, IND.

Dr. R. Sujithra @ Kanmani
Assistant Professor, School of Computer Science and Engineering, VIT University, Chennai, TN, IND.

ARTICLE INFO

Received: 02-02-2024
Received in revised form: 18-03-2024
Accepted: 20-03-2024
Available online: 30-03-2024

Keywords:

Disease Prediction;
Machine Learning;
Random Forest Algorithm;
Supervised Models.

ABSTRACT

The rise of computer-based innovations in the medical sector has led in electronic data acquiring. Because of the abundance of information readily available medical practitioners have the challenge of acknowledging indications and identifying diseases at an early stage. A wrong diagnosis is a major cause of inadequate therapy and failure to diagnose a serious illness in medicine. This paper assesses person's symptoms for disease prediction. In this paper we took input of three disease symptoms and evaluated them to give the disease as an output. Naive Bayes Classifier, Logistic Regression, K-Nearest neighbour (KNN), Support Vector Machine (SVM) and Random Forest Algorithm have been implemented in this paper. Our paper focuses on prediction of best accuracy model and also on the technique of splitting the dataset which will give us a better accuracy.

DOI: https://doi.org/10.54368/qijirse.3.1.0005

1.0 INTRODUCTION

Machine learning has become a crucial instrument in healthcare. It is a subset of artificial intelligence, that forecasts utilizing data instead of explicit programming. It has altered areas like as finance, marketing, and entertainment. Machine learning algorithms in healthcare can detect illnesses based on symptoms with remarkable accuracy and speed, improving patient outcomes and reducing the cost of healthcare. By utilizing huge amounts of healthcare data, machine learning is altering disease prediction in the healthcare industry. This technology not only improves illness prediction

★ *Corresponding author's e-mail:* elakiya.e@vit.ac.in (Dr. E. Elakiya)

accuracy and speed but also enables individuals to take control of their own health. Machine learning algorithms can save lives and save healthcare costs by using massive volumes of data, making it an appropriate application of information technology in the medical sector.

An extensive procedure encompassing patient history, physical examinations, and laboratory tests is needed for disease diagnosis. Conventional methods are tedious, expensive, and can sometimes be wrong. The global healthcare system is challenged with difficulties such as an aging population and an increasing load of chronic ailments, which require more efficient and precise diagnostic technology. Machine learning, which can analyse vast data, identify patterns, and predict results has the potential to overcome these challenges and alter illness diagnosis.

Early disease detection and diagnosis rely primarily on symptoms. Human error and variability can occur when medical professionals rely on clinical expertise. Machine learning can detect hidden patterns in an array of symptoms and data. Machine learning provides a more accurate and objective assessment of an individual's health by training models on large datasets. The availability of data is critical to the efficacy of machine learning in illness prediction, particularly from the healthcare industry's growing volume and variety of data sources, like EHRs, wearable devices, and patient-generated data through mobile apps, which can be used to train models and make evidence-based diagnostics. EHRs have become essential in modern healthcare as they offer a significant supply of data for machine learning. Algorithms can uncover subtle correlations between symptoms and diseases by analysing EHRs, resulting in early and accurate disease prediction and, ultimately, improved patient outcomes. It may also identify high-risk patients for preventative interventions, altering the focus of healthcare from reactive to proactive, saving lives, and lowering costs.

Machine learning-based illness prediction has the ability to alter individualized therapy by taking into account the individual characteristics of each patient, including symptoms, medical history, genetics, and lifestyle variables. This customized approach reduces misdiagnosis and ensures that treatments are better suited for individual patients, resulting in more effective and efficient healthcare.

Ethical issues are critical in machine learning, notably in disease prediction. Data privacy, security, and proper handling of patient information are crucial considerations. Machine learning algorithms, that require understanding and addressing, may have biases based on provided data. Improving the transparency and interpretability of these models is essential for providing equitable care and making anticipations that are comprehensible and acceptable to doctors and patients.

2.0 LITERATURE SURVEY

Gomathi *et al.,* (2022) in their work, tried to establish a machine learning-based disease prediction framework based on patient-reported symptoms with the aim to foster early disease identification, minimize doctor-patient interaction, and enable remote healthcare access. To evaluate patient symptoms and predict diseases, the system uses algorithms such as decision tree, random forest, SVM, and the Nave-Bayes classifier. The system has an accuracy rate of 93-95%, which can assist in future therapies and healthcare management. However, drawbacks include fears regarding algorithm bias, interpretability issues, and limited depth due to its concentration on symptoms compared to other factors. Kumar *et al.,* (2022) studied the potential of supervised machine-learning approaches to enhance early detection of high-risk diseases and address people's reluctance to seek medical care is examined in this study. The system examines user symptoms and predicts diseases employing a Naive Bayes Classifier, with the highest prediction accuracy reached for most diseases at 78.6%. For bigger model accuracy, the research suggests using complex machine learning

algorithms and expanding datasets to diverse groups. Integrating medical records into the system may minimize the strain on medical personnel while also improving the entire healthcare process. Data variety, model explanation, and a lack of real-world implementation details or clinical trials constitute some of the drawbacks. The aim of the research by Dahiwade *et al.,* (2019) is to predict diseases according to lifestyle and checkup statistics. It employs approaches to classify data and estimate associated dangers, especially K-Nearest Neighbour (KNN) and Covnets (CNN). In terms of accuracy, CNN surpasses KNN, with an accuracy rate of 84.5%. The process is efficient and low-cost, having the potential to drastically reduce the time required for accurate predictions. However, the study lacks information on disease symptom dataset sources, clinical validation, real-world testing, integration into healthcare practice, and the proposed system's scalability for managing many patients and their data. Rasheed and Glob, (2022) explored ML algorithms- SVM, KNN, Decision Trees, and Naive Bayes over the last three years, focusing on Naive Bayes, which obtained an 84% prediction accuracy rate for disease prediction. However, it lacks extensive information on data sources and quality, in addition to algorithm comparison. The paper highlights Naive Bayes' advantage without providing an exhaustive contrast with other machine learning approaches. A study by Kohli and Arora (2018) examines early disease recognition by comparing three disease databases: cardiovascular, breast cancer, and diabetes. For feature selection, the research utilized backward modelling with a p-value test. The findings show that multiple disease classifications helped obtain a better forecast accuracy, including 98.57% for Breast Cancer detection with AdaBoost classifier. 85.71% accuracy for Diabetes prediction with SVM, and 87.1% accuracy for cardiac disease prediction with Logistic Regression approach. The study also indicates prospects for data processing step automation and a pipeline structure for data pretreatment. However, the study lacks data on the disease dataset sources, clinical validation, and algorithm comparison, which could restrict the study's generalizability. Anishfathima *et al.,* (2022) predicted the likelihood of heart disease using methods based on machine learning including Random Forest, Support Vector Machine (SVM), decision tree, and logistic regression. The research does a comparative analysis to assess the efficacy of multiple methods in predicting cardiovascular diseases. SVM proved to be the most effective approach, having a 94% accuracy rate. due to the high mortality rate related to these diseases, the study highlights the need for accurate and successful cardiovascular disease prediction. Deep learning could be utilized to improve patient survival and well-being in the future, in addition to merging two distinct algorithms to predict abnormalities in the heart. The study, however, lacks information about the exact dataset and data sources used, and it fails to expand into clinical validation or practical applications. In several countries, increasing costs for healthcare are caused by a spike in the number of individuals and diseases in the medical system. For effective treatment, accurate disease prognosis is crucial. The study by Gupta and Gupta (2022) seeks to employ four distinct algorithms to forecast diseases based on patient symptoms, having an accuracy rate of 92-95%. This strategy reduces the need for hospital outpatient treatments and reduces the workload of medical staff. For the participation of users, an interactive interface is established. The research highlights the ease and low cost of algorithmic disease prediction providing suitable information and emphasizing the importance of accurate disease prediction in treatment.

In their study Anandajayam *et al.,* (2019) emphasizes the importance of evaluating medical data and predicting accuracy. The most accurate method is the Recurrent Neural Network (RNN), which surpasses alternatives like Naive Baye, KNN, Decision Trees, and SVM. Chronic diseases are a major issue, creating fatalities and causing a significant portion of patients' medical expenses. The article highlights the need to research these diseases and compares estimation algorithms, with RNN giving the best accuracy of 97.62% due to its feed-forward loop mechanism. Future studies may include using MRI scans and X-rays to boost forecast accuracy. The work, however, includes no

mention of clinical validation or real-world use of the methods. Sharma *et al.,* (2022) with the help of supervised machine learning algorithms have also demonstrated promise in predicting the risk of cardiovascular disease, COVID-19, and diabetes. For prediction, logistic regression is used, and it yields good accuracy across all datasets. The prediction model has effects on disease screening in real time. The article provides no mention of clinical validation or application of the methods. Despite this, the study emphasizes the potential of machine learning algorithms in improving disease diagnosis, especially in early detection of heart disease, COVID-19, and diabetes. Kanamarlapudi *et al.,* (2023) research project focuses on disease prediction based on symptoms, with users reporting their symptoms as inputs. The study relates many algorithms to solve enormous data handling problems and chooses the most accurate method for the problem statement. The main objective of the project is to predict diseases based on user inputs, with the disease providing the output. To ensure reliability, the study underlines the value of accuracy in disease prediction and the need for model review employing diverse methods. The goal of the research by Mittal and Sharma (2023) investigate the potential of machine learning in improving healthcare outcomes through assessing medical data additionally identifying minor changes that could indicate the existence of a certain disease. For identifying the best ML algorithm for prediction of disease, the study assesses approaches that include the decision tree algorithm and the Nave Bayes algorithm. The suggested method examines symptoms provided by user and displays condition's probability, using multiple ML models to evaluate their effectiveness in disease prediction. The paper additionally investigates the ethical implications of machine learning in healthcare and its significance in ethical usage. The approach based on decision trees was discovered to be the most accurate among high-level algorithms, with potential applications in early disease identification, such as cancer detection. The objective of the research by Çolak *et al.,* (2022) is to establish a high-accuracy disease prediction model employing several machine learning techniques. It analyses various algorithms utilizing a dataset comprising 133 symptoms, 42 disease groups, and 306 patient records. The results show that XGBoost surpasses other algorithms, having an accuracy score of 99.01%, making it suitable for tabular health data analysis. This research contributes to discipline by gaining the highest accuracy score for disease prediction reported in the literature. The small dataset, limited disease coverage, and restricted research setting are all limitations. The dataset should be extended in the future, and the system's performance in real-world healthcare scenarios needs to be assessed. When employing AI in healthcare, privacy and ethical issues must also be handled. Research by Pak and Shin (2014) focuses at the impact of external factors on the development of Type 2 diabetes, which has become an important issue in bioinformatics research. It intends to offer an analytical model for disease prediction through recognizing key environmental factors. Using the SMOTE method, the study resolves data imbalance in training data. The model uses SVM classifiers and training data from the Korean National Institute of Health's Ansan/Ansung Cohort 2 dataset. The results suggest that SVM classifiers based on meticulously selected environmental factors may surpass prediction models based only on genetic features in expecting Type 2 diabetes risk. The study highlights the potential efficiency of incorporating genetic and environmental variables in a cohesive approach to assess future disease risk. The research by Tang *et al.,* (2021) covers the under-diagnosis of Chronic Obstructive Pulmonary Disease (COPD) as an outcome from spirometry test underutilization. It examines the challenges in identifying the apt risk threshold and proposes the need for clinical prediction models that do not rely on spirometry. The study utilized neural network and logistic models to predict individuals' COPD risk projections, exploring with different brink levels. The study highlights the importance of evaluating the benefits of real positives against potential drawbacks of false positives. Continuous risk estimates are seen to be more accurate than risk group categorization. According to the results, clinical prediction models can be helpful for clinical decision-making,

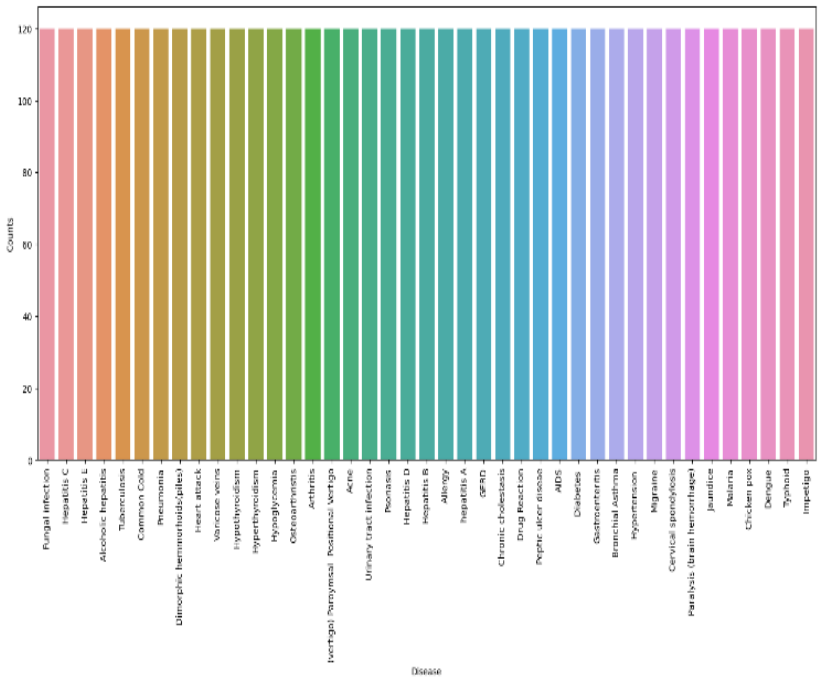
especially if clinical examination results are missing or unreported. However, it may not provide exhaustive details. In their paper, Huang *et al.,* (2022) deals with the assessment of schizophrenic patients on the basis of the impairments in their thoughts, language and communication. The standard method of assessments includes multiple sessions with mental health professionals to evaluate patients based on the PANSS and TLC scales. This paper aims to develop a machine learning model to evaluate patients based on semantic, syntactic and acoustic features of speech. This model will also be able to distinguish and analyse the stages of schizophrenia in positive as well as negative forms. The speech coherence of the paper is analysed to determine the stage of schizophrenia. The model proposed will target native Chinese speakers and incorporates models suitable for the linear representation of languages. The study by Le (2015) focuses on implementing computational approaches to predict disease-associated microRNAs (miRNAs), highlighting the importance of miRNAs in human diseases. It enhances the efficacy of network-based approaches including the Random Walk with Restart (RWR) approach by introducing RWRHMDA for ailment-associated miRNA prediction. The RWRHMDA algorithm depends on the RWRH approach, that was developed to analyse disease genes on some heterogeneous network of genes and symptoms of disease. The study evaluates RWRHMDA and RWRMDA performance on several disease phenotypes and finds that RWRHMDA surpasses RWRMDA across various parameter settings. RWRHMDA was employed in the investigation to discover eight novel miRNAs linked to Alzheimer's disease, highlighting their value in revealing new disease-associated miRNAs. However, the study might lack specifics on parameter settings or the real application itself. The study by Kanchan and Kishor (2016) seeks the high mortality rate from heart disease and diabetes, highlighting the underutilization of healthcare data for improved decision-making. It forecasts heart disease and diabetes utilizing techniques from machine learning such as SVM, Naive Baye, and Decision Tree, with and without PCA attribute reduction. After attribute reduction using PCA, the study demonstrates that SVM outperforms other algorithms for heart disease prediction. The research suggests developing a graphical user interface (GUI) system which employs SVM to predict cardiovascular illness. The study highlights the importance of ROC Area as an indicator of performance in diabetes prediction. The study, however, needs more a thorough examination of characteristics, limitations in data quality and availability, and an in-depth assessment of ethical and privacy concerns related to healthcare data mining. In their study, Dhekshagna *et al.,* (2023) uses techniques based on machine learning such as Nave Baye, Decision Tree, KNN, Random Forest and Logistic Regression to predict illnesses in the healthcare sector. The accuracy and F1_score measures are employed to evaluate the model's performance. In terms of reliability and accuracy, the Random Forest model surpasses other algorithms, while the Decision Tree and SVM models are effective. As a result of its higher accuracy, Random Forest is the apparent winner. This study enhances disease prediction abilities, providing helpful data for swift and accurate diagnosis. The study, however, lacks specifics regarding the dataset, its size, and potential limitations. It might benefit from delving into the larger ramifications of using machine learning in healthcare, such as ethical concerns and data privacy concerns. A convolutional neural network (CNN)- dependant unimodal disease risk prediction algorithm (CNN-UDRP) for heart disease prediction is offered in the article by Ambekar and Phalnikar (2018). The approach surpasses traditional methods for ML such as Naive Baye and KNN, with an accuracy of over 65 percent. The authors define potential advantages for disease prediction, such as reduced time consumption and cheaper expenses. The CNN-UDRP algorithm is an intriguing approach to disease risk prediction, surpassing conventional methods. The research, however, primarily limited to heart disease prediction and failed to evaluate its effectiveness in additional diseases. Additional research is needed to gauge the algorithm's generalizability and examine its application to a wider range of medical diseases. The objective of the initiative by Swarupa *et al.,* (2021) build an automated disease

prediction system that uses the Random Forest machine learning algorithm to help medical professionals in predicting and diagnosing diseases as early as feasible. The system should handle an extensive variety of symptoms, present a list of most likely diseases assessed by likelihood, and provide detailed explanations for each prognosis. The Python programming language and the Tkinter interface will be utilized for developing the system. The system's categorization accuracy, user interface, and explanations will be assessed. The subsequent phases are essential: data collection, preprocessing, feature selection, model training, evaluation, application, and monitoring. The system will be constructed based on a well-structured dataset and trained with the Random Forest technique. Based on the anticipated results, the system will provide accurate predictions and suggest further medical consultations or self-care actions.

3.0 METHODOLOGY

Examining the display of different AI computations is essential, and creating a tool to consider different AI machines in Python using scikit-learn will be useful. This test may be used as a framework for your AI problems and provide more information to consider. There will be a backup capacity for every model. You will get an understanding of how each model may match undetectable numbers by applying re-choice strategies such as cross-approval. It should have the option to utilize layouts to choose a couple of the best formats from the layout you have made. At the point when new information is free, it is essential to comprehend the data involving various advances to see the data in various ways. One thought is to pick a model. You should utilize a few unique strategies to see the similar components of your AI calculation to choose a couple to finish. The method for doing this is to utilize an assortment of procedures to show accuracy, variance, and different components of the plan dissemination.

3.1 Description of Dataset



The pandas library is used to load the dataset, and the null column is removed. The dataset is spotless, with no null values and just 0's and 1's as features. A bar plot is used to determine whether the target column is balanced. There are 120 samples for each disease in the dataset, thus no more

balancing is required. However, the target column, prognosis, is of the object datatype, making it unsuitable for training machine learning models. A label encoder is used to transform the prognostic column to numerical datatype, assigning a unique index to each label. If there are n labels, the labels' numbers will range from 0 to n-1. It's time to split the data to train and test the model now that we've cleaned it up by removing Null values and converted the labels to numerical representation. We will divide the data in a 70:30 ratio and 80:20 ratio, which means that 80% of the data will be used to train the model and 20% will be used to evaluate the models' performance.

3.2 Machine Learning Models

3.2.1 Random Forest Algorithm

An ensemble approach called Random Forest makes use of many decision trees to make predictions and combines their outputs to achieve high accuracy and robustness. For each tree, a random portion of the training data and a random subset of the characteristics are chosen to create these trees. The Random Forest uses a technique called bootstrapping to create multiple datasets from the original training data. Each decision tree is trained on one of these bootstrapped datasets. This helps reduce overfitting and improves generalization. Our model's random forest technique has an 89% accuracy rate when applied to test data.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (p_i)^2$$

3.2.2 K-Nearest Neighbours

For classification applications, the K-Nearest Neighbours Classifier is a simple machine learning model to apply. It works on the assumption that data points in a feature space that are similar belong to the same class. KNN is an instance-based or lazy learning method, implying that it does not directly build a model during training, but instead saves the full training dataset in memory for future comparisons. The "K" in KNN refers to the number of nearest neighbours taken into account for classification. When generating predictions, the algorithm uses a distance metric, such as Euclidean distance, the K training data points that are closest to the input data point should be found. The class designation is then assigned based on the majority vote of these neighbours. This easy-to-understand methodology makes KNN simple to use. The accuracy for K-Nearest Neighbours on test data in our model is 89.33%.

Common Distance Metrics: Euclidean distance (continuous distribution):

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

3.2.3 Support Vector Machine

One flexible machine learning technique that is well known for its ability to handle binary and multiclass classification issues is the Support Vector Machine (SVM) Classifier. The objective is to identify the most effective decision boundary, known as a hyperplane, that maximizes the margin between various categories in the data, enabling improved generalization and robustness. The ability of SVM to handle both linear and non-linear data separations, using linear kernel functions for linear problems and polynomial and radial basis functions for non-linear separations, differentiates it. This

modification allows SVM to capture challenging decision boundaries which were formerly not linearly separable in the feature space. SVM is well suited for challenges with a large number of features and has been widely used in a variety of fields that include text classification and image classification. The accuracy for Support Vector Machine Classifier in our model is 91.333%.

3.2.4 Gaussian Naïve Bayes

The machine learning model uses Gaussian Naive Bayes as its foundation. For categorization jobs, it is particularly created. It is designed for handling both binary and multiclass classification problems. It calculates the probabilities for each class and chooses the one with the highest probability for binary classification. The Bayes theorem is used by the probabilistic classifier Gaussian Nave Bayes to determine the likelihood that a certain occurrence falls into a particular class. With test data in our model, the Gaussian Naïve Bayes method has an accuracy of 86%. Measurements define the concept of Bayes as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

3.2.5 Logistic Regression

One popular machine learning approach for binary and multiclass classification problems is logistic regression. It uses the logistic function to transform input features into a positive class probability. It is a parametric model that estimates parameters during training. Key features include regularization methods, class-weight options, and multi-class strategies. It can handle both dense and sparse input data and is popular in fields like healthcare, finance, and natural language processing. The accuracy for Logistic Regression algorithm on test data in our model is 90%.

If p is the probability, then, the logit function for p is defined as:

$$\text{Logit}(p) = \text{In}\left(\frac{p}{1-p}\right) \text{ for } p \in (0,1)$$

3.3 Confusion Matrix, Table and Graphs

Table 1

80-20 Train Data

Parameters	Precision	Recall	F1 Score
SVM Classifier	89.97	90.23	90.10
Naïve Bayes Classifier	84.97	84.48	84.73
Random Forest	100.00	100.00	100.00
Logistic Regression	87.14	87.64	87.39
K Nearest Neighbours	90.09	88.79	89.44

Table 2

80-20 Test Data

Parameters	Precision	Recall	F1 Score
SVM Classifier	92.21	91.03	91.61
Naïve Bayes Classifier	88.51	83.97	86.18
Random Forest	89.68	89.1	89.39

Logistic Regression	90.38	90.38	90.38
K Nearest Neighbours	92.47	86.54	89.40

Table 3

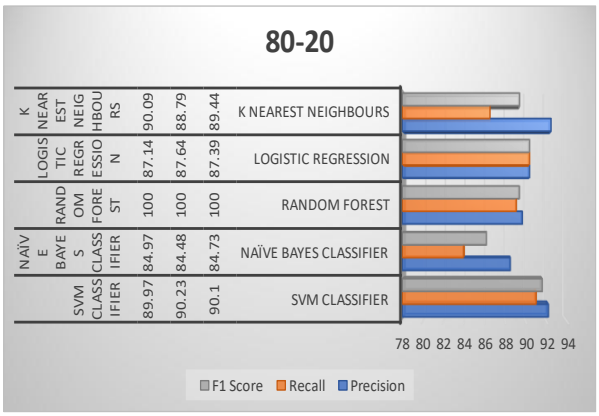
70-30 Train Data

Parameters	Precision	Recall	F1 Score
SVM Classifier	91.43	91.69	91.56
Naïve Bayes Classifier	86.26	84.53	85.38
Random Forest	100.00	100.00	100.00
Logistic Regression	87.92	89.68	88.79
K Nearest Neighbours	91.30	90.26	90.78

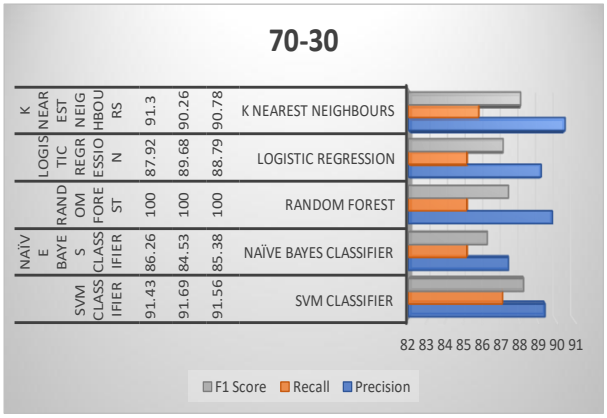
Table 4

70-30 Test Data

Parameters	Precision	Recall	F1 Score
SVM Classifier	89.40	87.10	88.24
Naïve Bayes Classifier	87.42	85.16	86.27
Random Forest	89.80	85.16	87.42
Logistic Regression	89.19	85.16	87.13
K Nearest Neighbours	90.48	85.81	88.08

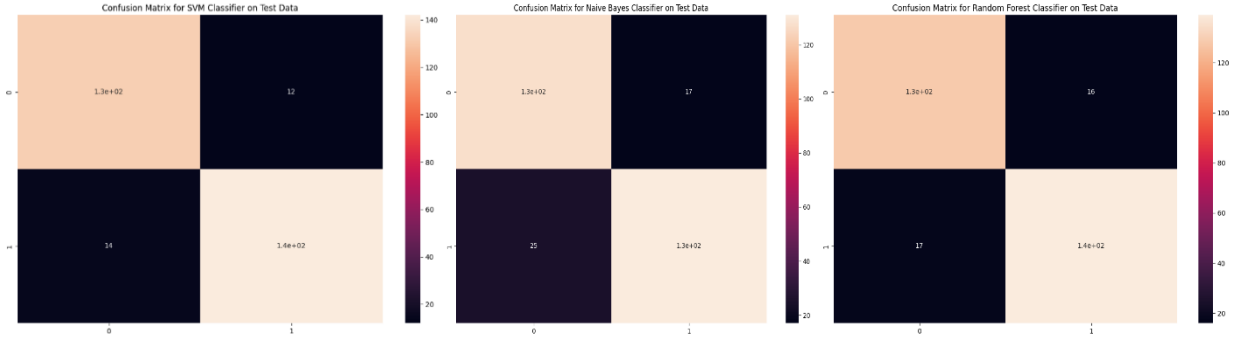


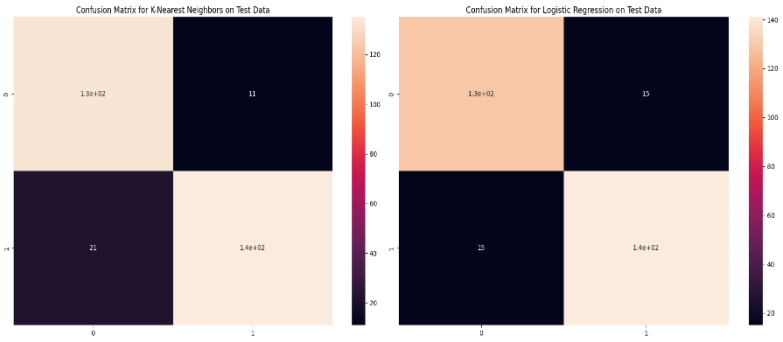
Graph 1



Graph 2

3.3.1 Confusion Matrix





4.0 RESULT

The accuracy comparison of the disease prediction for various machine learning algorithms are us follows.

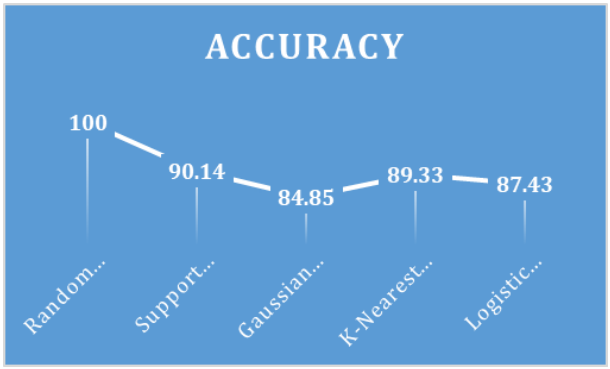
Table 5

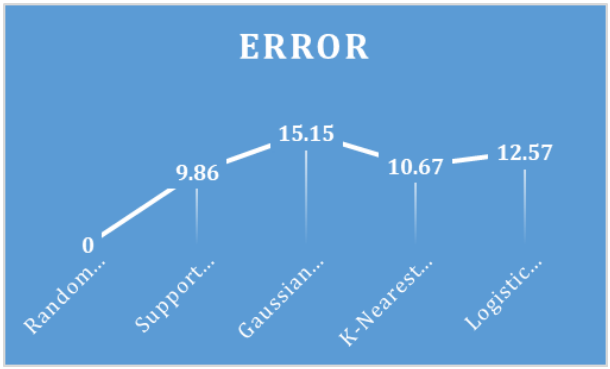
Comparison Analysis

Algorithm	Accuracy % on train data
Random Forest	100.00
Support Vector Machine	90.14
Gaussian Naïve Bayes	84.85
K-Nearest Neighbours	89.33
Logistic Regression	87.43

From the above table we conclude that all the algorithms are showing maximum accuracy from 84-100%. Random Forest model shows the highest accuracy of 100%

Algorithm	Error %
Random Forest	0.00
Support Vector Machine	9.86
Gaussian Naïve Bayes	15.15
K-Nearest Neighbours	10.67
Logistic Regression	12.57





5.0 CONCLUSION AND FUTURE WORK

The primary objective for this research is to predict illness based on symptoms provided by patients using machine learning algorithms that have been properly implemented.

Using the provided dataset, disease is automatically predicted with a maximum accuracy of 84–100%. In future work can be done using deep learning models. The above mentioned models have not been tested using actual real time data sets which can be extended further.

REFERENCES

Ambekar, S., & Phalnikar, R. (2018, August). Disease risk prediction by using convolutional neural network. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCUBEA.2018.8697423

Anandajayam, P., Aravindkumar, S., Arun, P., & Ajith, A. (2019, March). Prediction of chronic disease by machine learning. In *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICSCAN.2019.8878724

Anishfathima, B., Vikram, R., Sushaanthan, R. T., Vishnu, M. S., & Venumadhav, C. (2022, February). A Comparative Analysis on Classification Models to predict Cardio-vascular disease using Machine Learning Algorithms. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (pp. 259-264). IEEE. https://doi.org/10.1109/ICAIS53314.2022.9741831

Çolak, M., Sivri, T. T., Akman, N. P., Berkol, A., & Ekici, Y. (2022, September). A Study of Disease Prediction on Weighted Symptom Data Using Deep Learning and Machine Learning Algorithms. In *2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE)* (pp. 116-119). IEEE. https://doi.org/10.1109/ICTACSE50438.2022.10009857

Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1211-1215). IEEE. https://doi.org/10.1109/ICCMC.2019.8819782

Dhekshagna, G., Sushmitha, G., & Sairam, U. (2023, July). Geriatric Disease Prediction: A Study on Age-Based Disease Prognostication. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)* (pp. 983-986). IEEE. https://doi.org/10.1109/ICECAA58104.2023.10212172

Gomathi, R. M., Ajitha, P., Sivasangari, A., Anandhi, T., & Rani, V. N. (2022, April). Flawless Multi Perspective Vision for Prediction of Disease using Machine Learning Approach. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1190-1194). IEEE. https://doi.org/10.1109/ICOEI53556.2022.9776787

Gupta, A., & Gupta, M. K. (2022, April). Prediction of Diseases Using Different Machine Learning Approaches. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 712-717). IEEE. https://doi.org/10.1109/ICIEM54221.2022.9853132

Huang, Y. J., Lin, Y. T., Liu, C. C., Lee, L. E., Hung, S. H., Lo, J. K., & Fu, L. C. (2022). Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30*, 947-956. https://doi.org/10.1109/TNSRE.2022.3163777

Kanamarlapudi, S., Yakkala, V. S., Gayathri, B., Nusimala, K. V., Aravinth, S. S., & Srithar, S. (2023, February). Comparison and Analysis of Various Machine Learning Algorithms for Disease Prediction. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 246-250). IEEE. https://doi.org/10.1109/ICCMC56507.2023.10083509

Kanchan, B. D., & Kishor, M. M. (2016, December). Study of machine learning algorithms for special disease prediction using principal of component analysis. In *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)* (pp. 5-10). IEEE. https://doi.org/10.1109/ICGTSPICC.2016.7955260

Karunamurthy, A., Kulunthan, K., Dhivya, P. & Vickson, A. V. S. (2022). A Knowledge Discovery Based System Predicting Modelling for Heart Disease with Machine Learning. *Quing: International Journal of Innovative Research in Science and Engineering, 1*(1), 14-22. https://doi.org/10.54368/qijirse.1.1.0005

Kohli, P. S., & Arora, S. (2018, December). Application of machine learning in disease prediction. In *2018 4th International conference on computing communication and automation (ICCCA)* (pp. 1-4). IEEE. https://doi.org/10.1109/CCAA.2018.8777449

Kumar, R., Thakur, P., & Chauhan, S. (2022, May). Special disease prediction system using machine learning. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)* (Vol. 1, pp. 42-45). IEEE. https://doi.org/10.1109/COM-IT-CON54601.2022.9850843

Le, D. H. (2015, September). Disease phenotype similarity improves the prediction of novel disease-associated microRNAs. In *2015 2nd National Foundation for Science and Technology Development conference on information and computer science (NICS)* (pp. 76-81). IEEE. https://doi.org/10.1109/NICS.2015.7302226

Mittal, N., & Sharma, H. (2023, August). Automated Disease Prediction Using Machine Learning Technology. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 994-999). IEEE. https://doi.org/10.1109/ICIRCA57980.2023.10220645

Pak, M., & Shin, M. (2014, June). Developing disease risk prediction model based on environmental factors. In the *18th IEEE International Symposium on Consumer Electronics (ISCE 2014)* (pp. 1-2). IEEE. https://doi.org/10.1109/ISCE.2014.6884338

Ramalingam, A., Milan, A. J. & Mani, A. M. (2023). Artificial Intelligence and Machine Learning in Software Development. *Quing: International Journal of Innovative Research in Science and Engineering, 2*(2), 78-86. https://doi.org/10.54368/qijirse.2.2.0007

Rasheed, S. S., & Glob, I. H. (2022, December). Classifying and Prediction for Patient Disease Using Machine Learning Algorithms. In *2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA)* (pp. 196-200). IEEE. https://doi.org/10.1109/IT-ELA57378.2022.10107935

Sharma, A., Pathak, J., & Rajakumar, P. (2022, April). Disease Prediction using machine learning algorithms. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 995-999). IEEE. https://doi.org/10.1109/ICACITE53722.2022.9823744

Swarupa, A. N. V. K., Sree, V. H., Nookambika, S., Kishore, Y. K. S., & Teja, U. R. (2021, November). Disease prediction: smart disease prediction system using random forest algorithm. In *2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)* (pp. 48-51). IEEE. https://doi.org/10.1109/ICISSGT52025.2021.00021

Tang, Y., Zhang, L., Li, Q., & Zhu, S. (2021, July). Different thresholds in the prediction of chronic obstructive pulmonary disease using neural network and Logistic model. In *2021 International Conference on Public Health and Data Science (ICPHDS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICPHDS53608.2021.00008