WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

An Analysis and Comparison of various Algorithms using machine learning techniques in Identification and Prediction of Chronic Diseases in Human.

[1]Vineet Shrivastava, [2]Ashish Kumar Pandey, [3]Ashish Daksh, [4]Akash Kumar Singh, [5]Ansh Goel

[1, 2, 3, 4, 5] Computer Science & Engineering Department, Raj Kumar Goel Institute of Technology, Ghaziabad, UP, India.

ABSTRACT

Predicting disease from symptoms is an important area of research in health care, with important implications for early diagnosis, treatment effectiveness, and patient outcomes. Accurate disease prediction based on symptoms allows healthcare providers to intervene promptly, potentially preventing disease progression and improving the overall prognosis. While traditional methods of disease diagnosis rely heavily on clinical expertise and diagnostic tests, the rise of machine learning and artificial intelligence (AI) has opened up new possibilities for disease prediction based on symptom data. This analysis leads to simple diagnoses in the medical industry. and rapid treatment of patients Previous researchers have mainly focused on support vector machines (SVM), K-nearest neighbors (KNN), random forests, Naïve Bayes, and decision tree machine learning models for disease diagnosis with symptoms as parameters. Our diagnostic model serves as a virtual physician, enabling early disease detection and timely intervention. By incorporating feature engineering and standard scaling, we enhance the performance of our algorithms. This proposed model represents an important step towards automating healthcare processes, benefiting the entire industry.

INTRODUCTION

When anybody is presently suffering from an illness, they should visit a doctor, which consumes both time and money. It might also pose a challenge for the individual if they are far from medical professionals and facilities because the ailment cannot be recognized. Hence, if the aforementioned process could be automated through software that saves time and money, it might benefit the patient, ensuring a smoother procedure. All around the globe, chronic illnesses pose a significant challenge in the healthcare sector. As per medical reports, the mortality rate among individuals rises due to chronic conditions. The medical treatments prescribed for these ailments account for more than 70% of the patient's earnings. Therefore, it is crucial to reduce the risk factors contributing to the patient's death. Chronic diseases are a global healthcare challenge, significantly impacting mortality rates and consuming a substantial portion of patients' incomes. To mitigate the risk factors associated with these diseases, early detection is crucial. Artificial intelligence (AI) and machine learning offer promising avenues for improving disease diagnosis and prediction by developing algorithms that analyze symptoms and signs to identify diseases. Machine learning, a subset of AI, involves computer systems learning from data and experience to make predictions. By utilizing past data, machine learning algorithms can predict diseases based on patients' symptoms and medical histories. Despite decades of efforts, accurately predicting diseases remains a challenge in healthcare.

Efficiently addressing healthcare issues, including disease prediction, can be achieved through the application of machine learning technology. This research focuses on leveraging machine learning concepts to track and predict patient health, aiming to develop models that can predict diseases based on symptoms. Many developed nations, like India, are grappling with a variety of chronic illnesses, primarily cardiovascular disease and diabetes, which could have profound implications for global health, security, and the economy. The swift growth of cities and economies in today's world has led to a diverse range of lifestyles.

Chronic diseases have become a prevalent issue worldwide, impacting approximately one-third of the population in every country. The management of chronic diseases poses significant financial challenges and is particularly demanding for those aflicted. The healthcare industry gathers and analyzes extensive sets of chronic disease data, with data mining playing a crucial role in detecting diseases at an early stage. Among the most expensive diseases diagnosed are cardiovascular disease, diabetes, liver disease, Alzheimer's disease, and Parkinson's disease.

Overdiagnosis in healthcare is a prevalent issue that can result in unnecessary treatment and financial challenges. Factors contributing to misdiagnosis include unfamiliar symptoms, rare diseases, and conditions mistakenly overlooked.

Importance of Diseases from Symptoms The significance of predicting disease symptoms is its potential to transform healthcare allowing for proactive and treatment approaches. This seeks to contribute to the area by creating models of machine learning methods like trees and neural networks to forecast diseases based on symptoms Our objective is to the promise of machines in enhancing disease diagnosis management.

LITERATURE REVIEW

Akkem Yaganteeswarudu [1], the majority of disease prediction models are focused on analyzing a single disease at a time. However, the innovation of a multi-disease prediction model involves the utilization of machine learning and the Flask API to foresee various illnesses. This advancement eliminates the necessity for users to navigate through multiple models for disease predictions. Specifically, the system conducts evaluations for heart disease, breast cancer, diabetes, and diabetic retinopathy. The primary significance of this system lies in its ability to address a wider range of diseases for continuous patient monitoring and timely alerts to reduce fatality rates.

Dr. CK Gomathy, Mr. A. Rohith Naidu [2], The identification of illnesses is foreseen by the system based on symptoms by patients or any user. The user-entered symptoms are processed by the system as an input and generate disease probability as an output. Disease prediction utilizes a supervised machine learning method known as the Naive Bayes Classifier. This algorithm evaluates the likelihood of the disease occurring. Timely detection of diseases and patient well-being benefit from precise analysis of medical information, supported by biomedical and healthcare details. With the use of decision trees and linear regression techniques, illnesses such as diabetes, malaria, jaundice, dengue, and tuberculosis are predicted.

Melike Colak, Talya Tumer-sivri, Nergis Pervan-akman, Ali Berkol, and Yahya Ekici [3] established a fresh and trustworthy dataset for the prognosis of diseases through machine learning techniques. The dataset comprises medical information gathered from diverse origins, overseen by a medical specialist, encompassing 2006 patient records, 358 symptoms, and 141 illnesses. Various machine learning algorithms were applied in the research, such as boosting algorithms, to foresee diseases across different medical fields like diabetes, bronchial asthma, and COVID-19. The validations and comparison of the algorithms were conducted using cross-validation and multiple performance measurements. The study attained an impressive precision rate of 99.33% with a substantial number of ailments, marking a pioneering accomplishment in this field.

Swatik Paul, Pinku Ranjan, Somesh Kumar, Arun Kumar [4]. The primary algorithm utilized was the Random Forest Algorithm. The random forest algorithm was employed in the training of the model with a dataset that includes symptoms and the diseases they match. The reason behind selecting the random forest algorithm is its capacity to manage datasets comprising continuous variables similar to regression and categorical variables like classification. The random forest algorithm is known for delivering excellent outcomes in classification issues.

K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawanshi [5]. The disease prediction method relies on the symptoms and medical history of the patient, utilizing machine learning algorithms. The suggested model alters the information by giving importance based on rarity and trains it through Random Forest, LSTM, and SVM. By enhancing the accuracy and dependability of disease identification, the model aids in streamlining the healthcare sector.

PROPOSED METHODOLOGY

The upgraded and precise model for forecasting human diseases based on symptoms is offered by the suggested model.

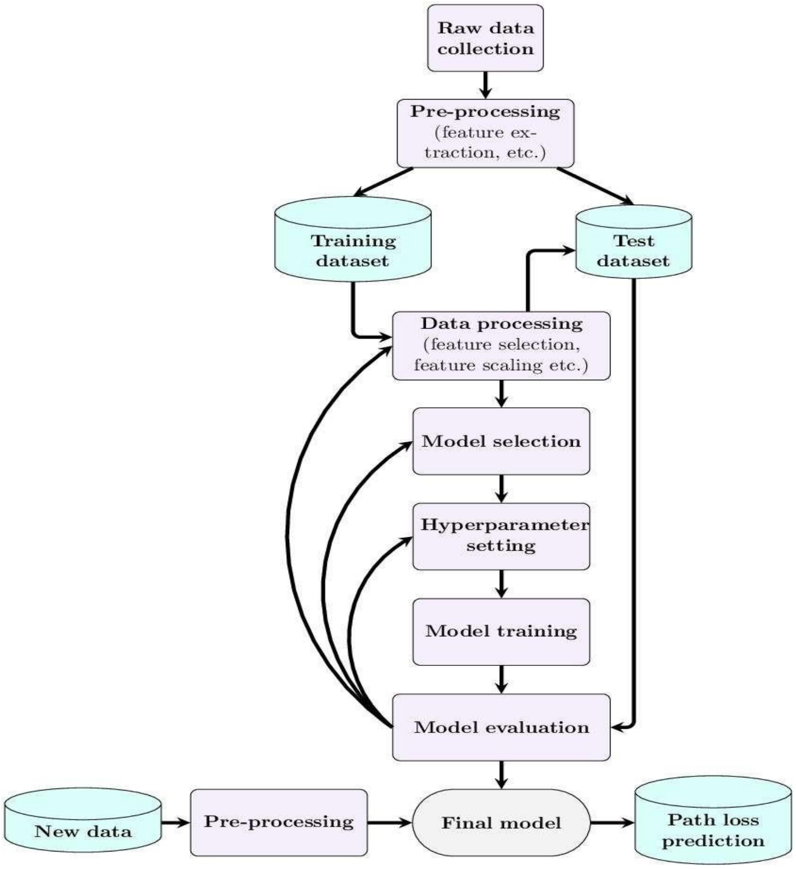

Fig.4.1 Flow diagram of proposed work

Data Collection: We obtained data sets from reliable sources that included information on symptoms and related disease outcomes. The data set includes a variety of symptoms, often associated with various diseases, and the corresponding disease diagnosis or label. The database has been carefully curated to ensure data quality and reliability.

Data processing: Before utilizing the dataset for model training, various preprocessing actions are taken to cleanse and organize the data. This involves managing missing values, encoding categorical variables, and standardizing numerical attributes. The dataset is then divided into training, validation, and test sets for assessing the model's performance.

Training data: The dataset training purposes are utilized in teaching machine learning. It comprises a collection of input features (including patient demographics, medical history, laboratory findings, etc.) and linked target indicators (showing whether the disease is identified or not). Within our framework, the training data amounts to 80%.

Test data: The test data set is used to evaluate the performance of the trained model. It is separate from the training data and contains unseen samples from the model that were not exposed during training. In our model, the test data is 20%.

Feature Selection: Using feature selection techniques to identify the most crucial features for predicting diseases. Employing methods like correlation analysis, feature importance, and domain knowledge to choose informative features for predicting disease symptoms. This aids in reducing database dimensions and enhancing model efficiency.

Model selection: We tested machine learning algorithms in the creation of our disease prediction models. Random Forests, Support Vector Machines, and Naive Bayes algorithms were among those considered. Hyperparameters for each algorithm were determined using methods like peering and cross-validation to enhance performance. Ultimately, after achieving a 100% accuracy rate across all algorithms tested, we opted to utilize a random forest classifier for model construction.

Naïve Bayes represents an artistic classifier grounded in principles of Bayes theorem, where it is independent among attributes. its straightforward nature finds broad application in tasks text categorization and detection. The process assesses the likelihood of a class for a specific by multiplying the prior of each class by the probability of the attributes with the class, which is then normalized to yield. Ultimately, the process identifies the class with the highest probability as the anticipated class for the. Naïvees stands out for its efficiency and the minimal data it demands owing to its basic parameter estimation. Nonetheless, the assumption of attribute independence does not always hold which has the potential to impact it. Despite this limitation, the simplicity, scalability, and efficacy of Nave Bayes across applications make it a preferred option for classification exercises. Moreover, Bayes' Theorem is articulated through the following equation.

$P (A/ B) = (P (B/ A) * P (A))/ P (B)$

Support Vector Machine (SVM): A mighty supervised training scheme is utilized for categorization and regression quandaries. The methodology involves identifying the best hyperplane that separates the data points of distinct classes with the largest margin. The goal of SVM is to diminish classification errors by maximizing the margin function and demarcating the gap between the hyperplane and the closest data point (support vector). In cases of linearly separable data, SVM pinpoints a linear hyperplane, while for non-linear data, it converts the input space to a feature space of higher dimensions employing a kernel function to accomplish linear segmentation. This transformation enables SVM to classify intricate data by pinpointing linear decision boundaries.

SVM is efficient in high-dimensional spaces, robust against redundancy, and suitable for both linear and nonlinear classification problems. However, the computational complexity of SVM can increase significantly with large databases, and it is important to choose the appropriate kernel and regularization parameters for optimal performance.

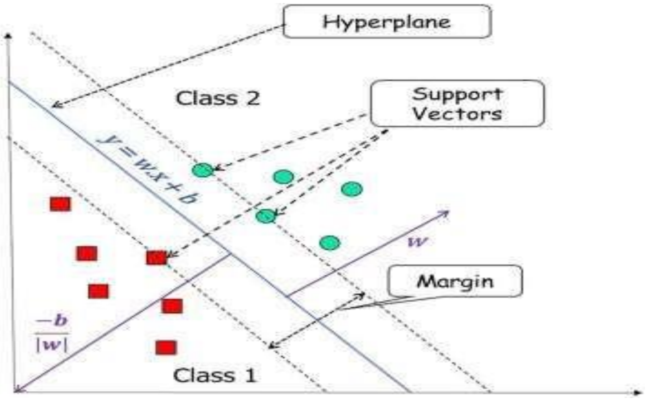

Fig. 4.2 Support Vector Machine

Random Forest: A multi-model ensemble classifier based on decision trees. It can be applied to regression in addition to classification. As the name implies, a random forest can be a classifier that employs normal sampling to increase the prediction accuracy of a given data set and comprises several decision trees for various subsets of that data set. Random forests use predictions from multiple decision trees and a high number of prediction votes to anticipate the final result, as opposed to depending just on one decision tree.

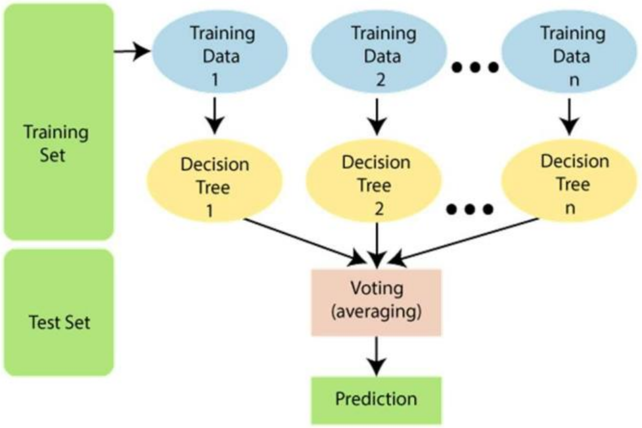

Fig.4.3 Random Forest

Model evaluation: We used metrics like accuracy, precision, recall, and F1 score to assess how well our models performed. Additionally, we evaluated the trade-off between sensitivity and specificity using methods like ROC curve analysis. To further illustrate our models' efficacy in illness prediction, we pitted their performance against baseline models.

(TP+TN)/(TP+FP+FN+TN) equals accuracy TP/(TP+FP) equals precision

Recall is equal to TP divided by (TP+FN)

* Precision * Recall / (Precision + Recall) = F1 score TP= True Positive

TN= True Negative

FP=False Positive FN= False Negative

Logistic regression: Logistic regression can be considered equivalent to using linear regression for situations where the target (or dependent) variable is discrete, i.e. not continuous. Theoretically, the response variable or label is binomial. A binomial response variable has two categories: yes/no, accept/not accept, default/not default, etc. Logistic regression is ideally suited for business analysis applications where the target variable is a binary decision (fail/pass, response/no response, etc.).

Decision trees: are one of the most powerful tools in supervised learning algorithms used for classification and regression problems. Each internal node represents a test on an attribute, each branch represents a test result, and each leaf node (terminal node) contains a class label. The training data is constructed by dividing the data into chunks based on attribute values until a stopping criterion, such as the maximum depth of the tree or the minimum number of samples required to split a node, is met.

During training, the Decision Tree Algorithm selects the best attributes to partition the data based on metrics such as entropy or impurity Gini, which measure the degree of impurity or randomness in the partition. The goal is to find the information partition or property that minimizes the impurity after partitioning.

RESULT AND DISCUSSION

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1**
Random Forest	**98.67%**	**98**	**98**	**98**
Decision Tree	**97.31%**	**98**	**97**	**97**
Support Vector Machine	**96.91%**	**97**	**97**	**96**
Naïve Bayes	**96.76%**	**97**	**96**	**96**
Logistic Regression	**96.56%**	**96**	**95**	**95**

Table 5.1 Result and Discussion

Here, we forecast the illness using five machine-learning algorithms. We obtained an accuracy of at least 97% for each of the five models. The Random Forest model produced the most accuracy, at 98.67%, while the Logistic Regression model produced the lowest accuracy, at 97.56%. In addition, the accuracy of the remaining models, DT, SVM, and Naïve Bayes, is correspondingly 98.31%, 97.91%, and 97.76%. Lastly, all of the models had 98% precision, recall, and F1 scores.

CONCLUSION

In this project, we developed a disease prediction system that can effectively predict people's diseases based on their symptoms. We use three machine learning algorithms to train and test our model on different disease databases: random forest, simplex, and vector machine. We compared model performance using accuracy scores and confusion matrices. Out of the three algorithms, we discovered that the random forest classifier has the best accuracy and the lowest error rate. We also put into practice a function that can receive symptoms as input and output a predicted illness. Our method can help with early disease detection and prevention, which will raise the standard of medical care.

The foundation of our system is the belief that disease databases contain actual cases and that symptoms are not independent. The category target column is converted to a numeric value using a label encoder. We used a 70:30 ratio to divide the data into training and test sets. To develop machine- learning algorithms and assess models, we make use of the Scikit-learning library. The final output is the last prediction, which we obtain by combining the predictions from the three models using a noise classifier. To locate more data and references on machine learning methods and illness prediction issues, we also used web search tools.

Some of the challenges and limitations of our system are a lack of adequate and reliable data for some rare diseases, missing or inaccurate symptoms, difficulty in treating many diseases with similar symptoms, and ethical and legal issues regarding personal use. health information. Some future areas and improvements of our system include: using more advanced and sophisticated machine learning algorithms such as deep learning and natural language processing, age, gender, lifestyle, medical history, etc. including more features and factors such as web or mobile applications that are user- friendly and interactive, ensuring privacy and security of data and predictions.

FUTURE WORK

Future studies in this research will focus on various areas to improve clinical practice and its clinical outcomes. First, combining characteristics other than symptoms, such as demographics, medical history, and lifestyle, could be explored to improve the accuracy and affordability of disease prediction. Continuous improvement and optimization of machine learning models by experimenting with different algorithms, feature engineering techniques, and hyperparameter modification methods will make them more accurate. In addition, the development of systems that integrate real-time data such as wearable devices, electronic health records, and mobile phones will be able to offer individual health measures to users. Efficacy studies in clinical environments are important in terms of evaluating the effectiveness of the application in real-life conditions and its applicability to physicians and patients. Additionally, expanding the app's functionality to cover more medical and clinical areas and investing in user interface (UI) and user experience (UX) improvements can make the app more intuitive and accessible to people of all skill levels. personal use information and health information. Longitudinal study data to assess disease progression, treatment, and patient outcomes over time can provide insight into individual treatment plans and eight influencers. Finally, working with doctors and hospitals to integrate medical records into their existing workflows and systems can facilitate communication and exchange of information regarding patients and physicians.

REFERENCES

1. Akkem Yaganteeswarudu(2020) "Multi Disease Prediction Model by using Machine Learning and Flask API "**,** 5th International Conference on Communication and Electronics Systems (ICCES).
2. Dr. CK Gomathy, Mr. A. Rohith Naidu(2021)," The prediction of disease using machine learning " International Journal of Scientific Research in Engineering and Management (IJSREM).
3. Melike colak ,Talya Tumer-sivri ,Nergis pervan-akman ,Ali Berkol and Yahya Ekici (2022)," A Study of Disease Prediction on Weighted Symptom Data Using Deep Learning and Machine Learning Algorithms " International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE).
4. Swatik Paul, Pinku Ranjan, Somesh Kumar, Arun Kumar (2022)," Disease predictor using random forest classifier " 2022 International Research for Advancement in Technology.
5. K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawanshi (2023)," Human Disease Prediction using Machine Learning Techniques and Reallife Parameters", International Journal of Engineering (IJE).
6. Farooqui, M. and Ahmad, D., (2020)"Disease prediction system using support vector machine and multilinear regression", International Journal of Innovative Research in Computer Science & Technology (IJIRCST).
7. Rathi, M. and Pareek, V., (2020)"Disease prediction tool: An integrated hybrid data mining approach for healthcare", IRACSTInternational Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN.
8. Paul, S., Ranjan, P., Kumar, S. and Kumar, A., (2022) "Disease predictor using random forest classifier", in International Conference for Advancement in Technology (ICONAT), IEEE.
9. Rahman, A.S., Shamrat, F.J.M., Tasnim, Z., Roy, J. and Hossain, S.A., (2019)"A comparative study on liver disease prediction using supervised machine learning algorithms", International Journal of Scientific & Technology Research.
10. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S. and Karve, A., (2019), "Disease prediction using machine learning", International Research Journal of Engineering and Technology (IRJET).
11. Nishant Yede, Ritik Koul, Chetn Harde, Kumar Gaurav and Prof. C.S. Pagar, "General Disease Prediction Based On Symptoms Provided By Patient", *Open Access International Journal Of Science & Engineering (OAIJSE)*, June 2021.
12. Rinkal Keniya, Aman Khakharia and Vruddhi Shah, "Disease Prediction From Various Symotoms Using Machine Learning", *Social Science Research Network(SSRN)*, October 2020.
13. Sneha Grampurohit and Chetan Sagamal, "Disease Prediction using Machine Learning Algorithms", *IEEE International Conference for Emerging Technology (INCET)*, August 2020.
14. Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach", *IEEE 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, August 2019.