International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Transforming Healthcare: Predicting Diseases using Machine Learning and AI

Pranit Bawane

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

With the rapid growth of artificial intelligence, machine learning (ML) has become a game-changer in healthcare, providing innovative ways to predict diseases and diagnose them early. Traditional diagnostic methods often depend on manual assessments, which can be slow, prone to errors, and reliant on expert interpretation. By integrating ML techniques into disease prediction systems, we can significantly boost accuracy, efficiency, and accessibility in medical diagnostics.

This study introduces a machine learning-based disease prediction system that utilizes patient medical data— like symptoms, demographic details, and clinical history—to estimate the likelihood of various diseases. We conduct a comparative analysis of several ML algorithms, including Decision Trees, Random Forest, Support Vector Machines (SVM), and Deep Learning models, to find the most effective approach. The dataset is refined through preprocessing techniques such as feature selection, data normalization, and addressing missing values to enhance model performance. We evaluate the models based on accuracy, precision, recall, and F1-score to ensure reliable predictions.

The experimental results show that ML-based disease prediction surpasses traditional diagnostic methods, providing better accuracy and quicker processing times. These findings highlight that machine learning can be vital in early disease detection, ultimately lowering mortality rates and improving patient care.

This research emphasizes the increasing significance of artificial intelligence in the healthcare field. Future efforts will aim to expand the dataset, incorporate deep learning architectures, and integrate real-time patient monitoring systems to further enhance disease prediction capabilities. By advancing ML-based diagnostic tools, this study contributes to the creation of intelligent healthcare solutions that improve early detection, treatment planning, and overall patient outcomes.

KEYWORDS: *Disease Prediction, Machine learning.*

I. INTRODUCTION

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience. The machine learning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient's symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying complete machine learning concepts to keep the track of patient's health.

ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services. To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm.

Human disease predication is a crucial part of human life. Early disease prediction of a human is an important step in the treatment of disease. Since the very beginning, a doctor has handled it almost exclusively. Thus, the healthcare industry thrives on innovation to make logistics efficient. Innovation is the heart of the medical industry. It is what drives new treatments, cures and therapies. Innovation is also what keeps the medical industry current and relevant. The scope of development in the medical industry is vast. There are many areas where innovation is needed to make progress. Some of these include developing new treatments for diseases, finding ways to improve patient care, and making medical procedures more efficient. In the current digital age, innovation in the medical industry can be achieved through the digitalization of medical. One of the most pressing issues in the medical industry is the workload on the doctors and the unaffordable consultation cost. This issue is highlighted mainly in the disease prediction with the symptoms of the patients as input. The current methodology of the medical industry consists of the patient visiting a generalist doctor and explaining to the doctor the conditions, and symptoms faced by the patient upon which the doctor infers possible diseases and then channels them to a specialist doctor. The logistics behind this methodology can be minimized with the help of a machine learning algorithm: Random Forest. This algorithm is used for classifying

multiple diseases based on symptoms and geographic locations. These locations help determine the results as the database assumes that for a particular location, there exist some symptoms that only occur at that location. Thus, unlike other models, this model concentrates.

II. EXISTING SYSTEM

The existing system predicts the chronic diseases which are for a particular region and for the particular community. Only particular diseases are predicted by this system. In this System, Big Data & CNN Algorithm is used for Disease risk prediction. For S type data, the system is using Machine Learning algorithm i.e. Knearest Neighbors, Decision Tree, Naïve Bayesian. The accuracy of the existing System is up to 94.8%. In the existing paper, they streamline machine learning algorithms for the effective prediction of chronic disease outbreak in disease-frequent communities. They experiment with the modified prediction models over real-life hospital data collected from central China. They propose a convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from the hospital.

III. METHODOLOGY

The proposed model is providing an enhanced and accurate model for predicting human diseases from the symptoms. The dataset from Kaggle is used, and the methods used to train the models are the Rainforest algorithm, LSTM algorithm and SVM algorithm to train our data. The working model will be as follows: 1. the human will enter his/her symptoms. 2. The symptoms will then be inputted into our model. 3. The model will then yield the possible disease. The novelty of the proposed work is that tweaking the Random forest model by using hyper parameters, improves the efficacy of the model. Hence, it is providing more accuracy.

To calculate performance evaluation in the experiment, first, we denote TP, TN, Fp and FNias true positive(the number of results correctly predicted as required), true negative (the number of results not required), false positive (the number of results incorrectly predicted as required), false negative(the number of results incorrectly predicted as not required)respectively. We can obtain four measurements: recall, precision, accuracy, and F1 measures as follows:

Accuracy-:

$$\frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative}$$

$$F1\text{-}Measure = \frac{2\times precision\times recall}{precision+recall}$$

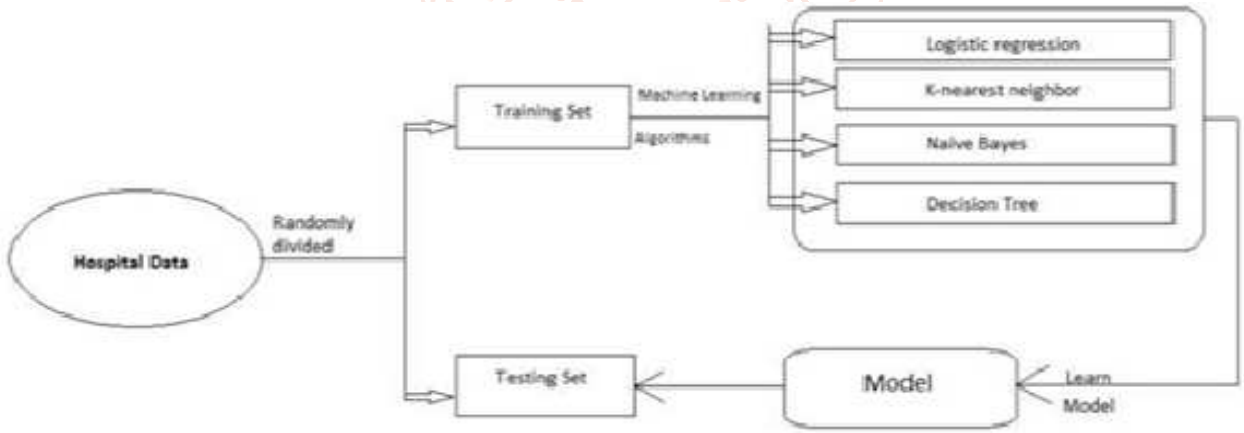

Fig 1. System Architecture

IV. RESULTS

A decision tree is a structure that can be used to divide up a large collection of records into successfully smaller sets of records by applying a sequence of simple decision tree. With each successive division, the members of the resulting sets become more and more similar to each other. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous (mutually exclusive) groups with respect to a particular target. The target variable is usually categorical and the decision tree is used either to:

- Calculate the probability that a given record belong to each of the category and,
- To classify the record by assigning it to the most likely class (or category). In this disease prediction system, decision tree divides the symptoms as per its category and reduces the dataset difficulty

GRAPHS:

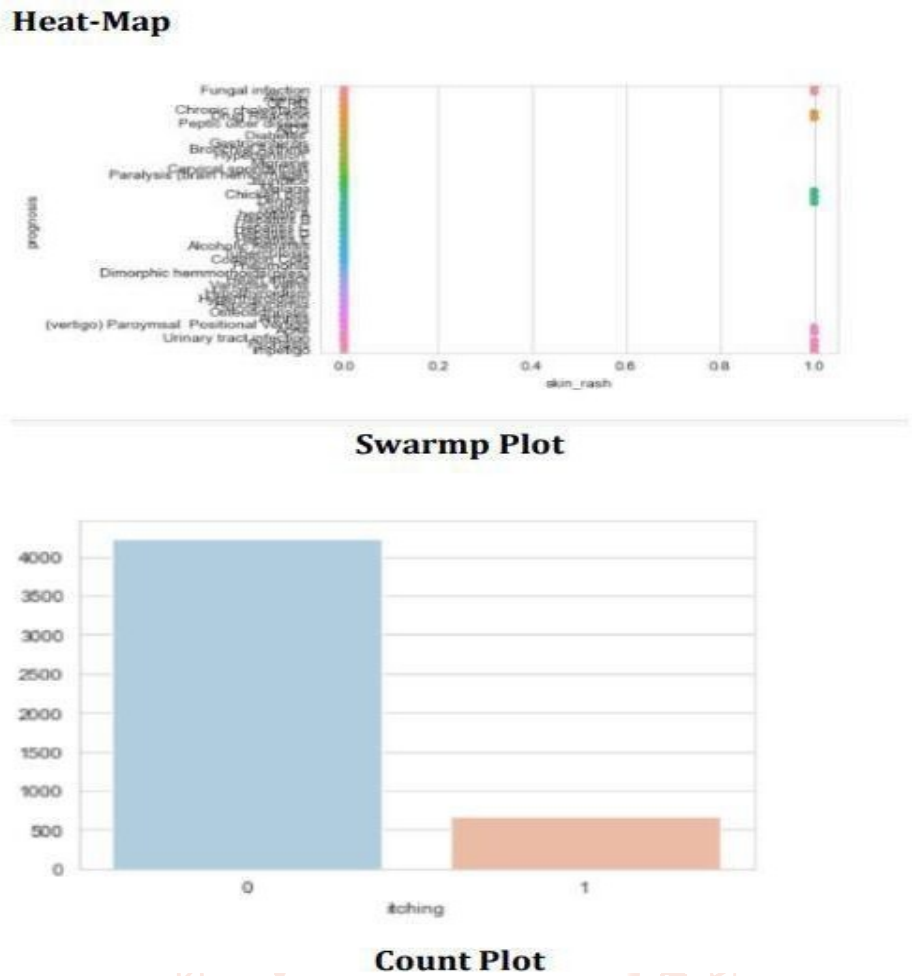

Fig 2: Graph (Prediction, Swarmp, Count)

Comparison of accuracy of algorithm.
Decision Tree 84.5%
Random Forest 98.95%
Naïve Bayes 89.4%
SVM 96.49%
KNN 71.28%

We found that the Support Vector Machine (SVM) algorithm is widely used (in 30 studies) followed by the Naïve Bayes algorithm (in 24 studies). However, the Random Forest algorithm showed relatively high accuracy. In the 40 studies in which it was used, RF showed the highest accuracy of 98.95%.

This was followed by SVM which included 96% of the accuracy considered.

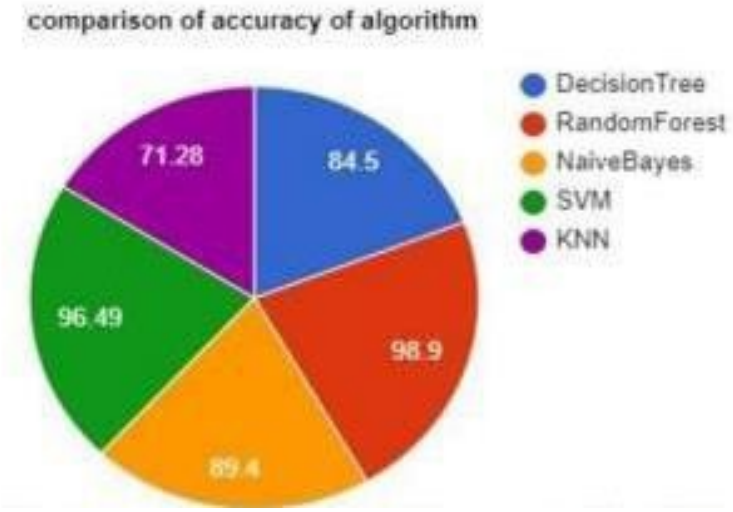

Fig 3. ML Model Accuracy in Healthcare Prediction

V. CONCLUSION

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease. Average prediction accuracy probability of 100% is obtained. Disease Predictor was successfully implemented using the grails framework. This system gives a user-friendly environment and easy to use.

As the system is based on the web application, the user can use this system from anywhere and at any time. In conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of the hospital data.

This systematic review aims to determine the performance, limitations, and future use of Software in health care. Findings may help inform future developers of Disease Predictability Software and promote personalized patient care. The program predicts Patient Diseases. Disease Prediction is done through User Symbols.In this System Decision tree, Unplanned Forest, the Naïve Bayes Algorithm is used to predict diseases. For the data format, the system uses the Machine learning algorithm Process Data on Database Data namely, Random Forest, Decision Tree, and NaiveBayes. System accuracy reaches 98.3%. Machine learning skills are designed to successfully predict outbreaks.

Table 1. Comparative Analysis

Ref.	Algorithm Used	Advantages	Limitation(s)	Accuracy
[17]	Naive Bayes Classifier	Highly Scalable	Only for independent features it works accurately	94.8%
[18]	Random forest, Decision tree, Naïve Bayes	Good accuracy for predicting disease	Model needs to be enhanced via ensamble model	90%
[15]	Weighted KNN	Smoother decision surface, less data dependency	Due the issue of over-fitting, model is not scalable	93.5%
[29]	SVM	Faster Execution, Less Space complexity	Not Suitable for Multi-parameter	76%
[30]	SVM	Faster Execution, Less Space complexity	Not Suitable for Multi-parameter	90%
[32]	Logistic Regression(LR)	It makes assumption about distribution	Over-Fitting issue is there. It requires less multi-collinearity	75%
Proposed Method	Random Forest	The dataset is suitable for Random Forest	Can be improved if time series dataset is provided	97%

VI. REFERENCES

[1] Zhou, S.-M., Fernandez-Gutierrez, F., Kennedy, J., Cooksey, R., Atkinson, M., Denaxas, S., Siebert, S., Dixon, W. G., O'Neill, T.W. and Choy, E., "Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis", PloS One, Vol. 11, No. 5, (2016), https://doi.org/10.1371/journal.pone.0154515

[2] Littell, C.L., "Innovation in medical technology: Reading the indicators", Health Affairs, Vol. 13, No. 3, (1994), 226-235. https://doi.org/10.1377/hlthaff.13.3.226 3. 4. 5. 6. 7. 8. 9.

[3] Milella, F., Minelli, E.A., Strozzi, F. and Croce, D., "Change and innovation in healthcare: Findings from literature", ClinicoEconomics and Outcomes Research, (2021), 395-408. doi:10.2147/CEOR.S301169.

[4] Rathi, M. and Pareek, V., "Disease prediction tool: An integrated hybrid data mining approach for healthcare", IRACST International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN, (2016), 2249-9555.

[5] Kelly, C.J. and Young, A.J., "Promoting innovation in healthcare", Future Healthcare Journal, Vol. 4, No. 2, (2017), 121. doi:10.7861/futurehosp.4-2-121.

[6] Mobeen, A., Shafiq, M., Aziz, M.H. and Mohsin, M.J., "Impact of workflow interruptions on baseline activities of the doctors working in the emergency department", BMJ Open Quality, Vol. 11, No. 3, (2022), e001813. doi:10.1136/bmjoq-2022 001813.

[7] Ahmed, S., Szabo, S. and Nilsen, K., "Catastrophic healthcare expenditure and impoverishment in tropical deltas: Evidence from the mekong delta region", International Journal for Equity in Health, Vol. 17, No. 1, (2018), 1-13. doi:10.1186/s12939-018-0757-5.

[8] Roberts, M.A. and Abery, B.H., "A person-centered approach to home and community-based services outcome measurement", Frontiers in rehabilitation Sciences, Vol. 4, (2023). doi:10.3389/fresc.2023.1056530

[9] Farooqui, M. and Ahmad, D., "Disease prediction system using support vector machine and multilinear regression", International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, (2020), 2347-5552.

[10] Olatunji, O.O., Adedeji, P.A., Akinlabi, S., Madushele, N., Ishola, F. and Aworinde, A.K., "Improving classification performance of skewed biomass data", in IOP Conference Series: Materials Science and Engineering, IOP Publishing. Vol. 1107, (2021), 012191. 11.

[11] Cao, J., Wang, M., Li, Y. and Zhang, Q., "Improved support vector machine classification algorithm based on adaptive feature weight updating in the hadoop cluster environment", PloS One, Vol. 14, No. 4, (2019), e0215136. https://doi.org/10.1371/journal.pone.0215136 12.

[12] Hamidi, H. and Daraee, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", International Journal of Engineering, Transactions B: Applications, Vol. 29, No. 7, (2016), 921-930. 13.

[13] Pisner, D.A. and Schnyer, D.M., Support vector machine, in Machine learning. 2020, Elsevier.101-121.

[14] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011; 12:2825-2830.