IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303276 |

Disease Prediction using Machine Learning

Dr. Nitish Das[1], Sailee Gayke[2], Nayan Patel[3], Shrawani Shinde[4]

Professor, MIT Art, Design & Technology University, Loni Kalbhor, Maharashtra, India[2]

Students, Department of Computer Engineering, MIT Art, Design & Technology University, Loni Kalbhor, Maharashtra, India[2,3,4]

ABSTRACT: The "Disease Prediction" method, which is concentrated on predictive modelling, it predicts the user's disease based on the symptoms that the user provides as input. The method examines the user's symptoms as input and returns the disease's likelihood as an output. Disease prediction is accomplished using the random forest classifier. Disease prediction of a human means predicting the probability of a patient's disease after examining the combinations of the patient's symptoms. Monitoring a patient's condition and health information at the initial examination can help doctors to treat a patient's condition effectively. This analysis in the medical industry would lead to a streamlined and expedited treatment of patients. The previous researchers have primarily emphasized machine learning models mainly Support Vector Machine (SVM), K-nearest neighbours (KNN), and RUSboost for the detection of diseases with the symptoms as parameters. However, the data used by the prior researchers for training the model is not transformed and the model is completely dependent on the symptoms, while their accuracy is poor. Nevertheless, there is a need to design a modified model for better accuracy and early prediction of human disease. The proposed model has improved the efficacy and accuracy model, by resolving the issue of the earlier researcher's models. The proposed model is using the medical dataset from Kaggle and transforms the data by assigning the weights based on their rarity. This dataset is then trained using a combination of machine learning algorithms: Random Forest, Long Short-Term Memory (LSTM), and SVM. Parallel to this, the history of the patient can be analyzed using LSTM Algorithm. SVM is then used to conclude, the possible disease. The proposed model has achieved better accuracy and reliability as compared to state-of-the-art methods. The proposed model is useful to contribute towards development in the automation of the healthcare industries.

KEYWORDS: Random Forest, Support Vector Machine, Symptoms, Disease Prediction, Adaboost, Machine Learning, K-nearest neighbours (KNN), etc.

I. INTRODUCTION

When anyone is currently afflicted with an illness, they must see a doctor, which is both time consuming and costly. It can also be difficult for the user if they are not near to doctors and hospitals because the illness cannot be identified. So, if the above procedure can be done using automated software that saves time and money, it could be better for the patient, making the process go more smoothly. There are other Heart Disease Prediction Systems that use data mining methods to analyse the patient's risk level. Disease Predictor is a web-based system that predicts a user's disease based on the symptoms they have. Data sets from various health related websites have been obtained for the Disease Prediction system. The consumer will be able to determine the likelihood of a disease based on the symptoms given using Disease Predictor. People are always curious to learn new things, particularly as the use of the internet grows every day. When an issue occurs, people often want to look it up on the internet. Hospitals and physicians have less access to the internet than the general public. When people are afflicted with an illness, they do not have many options. As a result, this system can be beneficial to people. Chronic illness is a disease that lasts a long time or takes a long time to heal, and many chronic diseases cannot be cured but can only be managed with daily treatments. India, like all other nations, is undergoing significant social and economic shifts, which is causing a fast rise in the frequency of cardiovascular disease.

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 [1]. Innovation is the heart of the medical industry. It is what drives new treatments, cures and therapies [2]. Innovation is also what keeps the medical industry current and relevant. The scope of development in the medical industry is vast [3, 4]. 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,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303276 |

and making medical procedures more efficient. In the current digital age, innovation in the medical industry can be achieved through the digitalization of medical processes [5]. One of the most pressing issues in the medical industry is the workload on the doctors [6] and the unaffordable consultation cost [7]. 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 [8]. The logistics behind this methodology can be minimized with the help of a machine learning algorithm: Random Forest [9]. 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.

II. RELATED WORK

As discussed in the introduction sections, some of the research papers include a plethora of models for predicting the disease that a patient may suffer, based on symptoms gathered from the patient. The models that are used often and have the best accuracy are as follows:

The Method proposed by Jianfang et al. [11] used Support Vector Machine (SVM) for the classification of diseases based on the symptoms. The SVM model is efficient for the prediction of diseases but requires more time to predict disease [12]. Also, a method is unable to increase the accuracy of the model. The approach has the drawback of classifying objects using a hyperplane, which is only partially effective [13, 14]. The hyperplane is accurate only for classifying sample data into 2 classes. But in the current scenario, the medical industry requires more than 2 classes (diseases) for the identification of symptoms corresponding to the disease.

The K-Nearest Neighbors (KNN) algorithm used by Keniya et al. [15]. They used this method by assigning the data point to the class that most of the K data points belong to, while it is sensitive to noisy and missing data. They have considered certain factors such as age group, symptoms and gender of the person to predict the disease. While considering these parameters lower accuracy on machine learning models is getting [15]. The KNN method is also used by Kashvi et al. [16]. They also have proven high accuracy in several cases such as diabetics and heart risk prediction. There is the issue of considering a small data size for the classification of diseases [16].

The method proposed by Pingale et al. [17] using Naïve Bayes method they are predicting limited diseases such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis They have not worked on a large dataset to predict large numbers of diseases [17]. Also, Gomathy, and Rohith Naidu [18] used Naïve Bayes method for disease prediction. By using this method they have developed a web application for disease prediction that is accessible from anywhere. The accuracy of the model depends on the data provided to the system. The issue of the suggested model is to develop software for disease prediction with a more accurate dataset to enhance the accuracy [18]. The method proposed by Chhogyal and Nayak [19] used Naïve Bayes classifier. They have obtained poor accuracy in disease prediction also they are not using the standard dataset for training [19].

The method proposed by Kumar et al. [20] used Rustboost Algorithm. RUSBoost is developed to address the issue of class imbalance [20]. However, the RUSBoost algorithm employs random under-sampling as a resampling method which can lead to the loss of crucial information. Therefore, this algorithm was not taken into account when training the data.

The above-mentioned approaches have discussed various machine-learning techniques for disease prediction. However, the author has not employed some issues such as efficiency, accuracy; the limited size of the data set used to train the model and considered limited symptoms to diagnose the disease. To overcome all these issues there is a need to propose a modified and accurate model for predicting human diseases. The detailed proposed model is described below section.

III. METHODOLOGY

Proposed 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.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303276 |

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.
In this work standard dataset is used for training and testing the model, author has tested multiple models including the models discussed under the section "Literature Review". With the conclusion to the experiment, the following combinations of methodologies are used in the proposed model:
This article aims to implement a robust machine-learning model that can efficiently predict the disease of a human, based on the symptoms that he/she possesses. Let us look into how we can approach this machine-learning problem:

Approach:

- **Gathering the Data:** Data preparation is the primary step for any machine learning problem. We will be using a dataset from Kaggle for this problem. This dataset consists of two CSV files one for training and one for testing. There is a total of 133 columns in the dataset out of which 132 columns represent the symptoms and the last column is the prognosis.
- **Cleaning the Data:** Cleaning is the most important step in a machine learning project. The quality of our data determines the quality of our machine-learning model. So it is always necessary to clean the data before feeding it to the model for training. In our dataset all the columns are numerical, the target column i.e. prognosis is a string type and is encoded to numerical form using a label encoder.
- **Model Building:** After gathering and cleaning the data, the data is ready and can be used to train a machine learning model. We will be using this cleaned data to train the Support Vector Classifier, Naive Bayes Classifier, and Random Forest Classifier. We will be using a confusion matrix to determine the quality of the models.
- **Inference:** After training the three models we will be predicting the disease for the input symptoms by combining the predictions of all three models. This makes our overall prediction more robust and accurate.

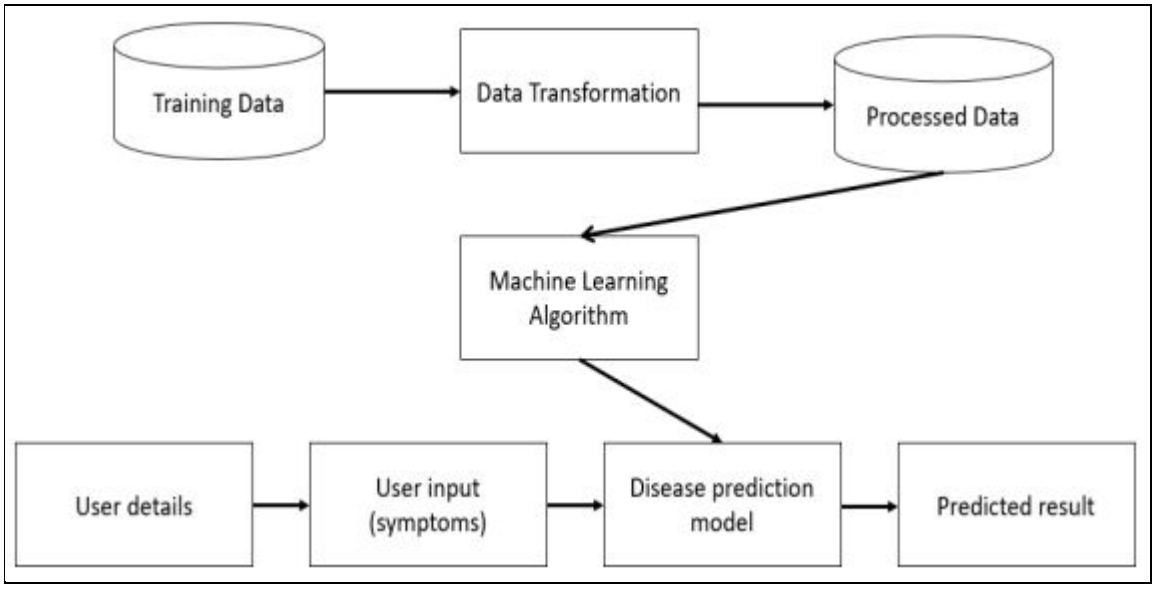

Fig.1: Proposed System Block Diagram

IV. COMPARATIVE ANALYSIS

In the comparative analysis of disease prediction using machine learning, we evaluated the performance of several prominent algorithms to ascertain their effectiveness in predicting diseases based on patient data. The dataset employed for this analysis comprised diverse features such as patient demographics, medical history, and lifestyle factors. Three machine learning models were selected for comparison: Support Vector Machines (SVM), Random Forests, and Logistic Regression.
In the methodology, the dataset underwent pre-processing, including handling missing values, encoding categorical variables, and scaling numerical features. A stratified split was employed to create training and testing sets, ensuring a representative distribution of disease instances across both subsets. Each model underwent training on the training set,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303276 |

with hyper-parameter tuning performed through cross-validation. The test set was then used to evaluate the models' performance based on key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

The results revealed nuanced differences among the models. Support Vector Machines exhibited commendable accuracy and precision, particularly for diseases with imbalanced class distributions. Random Forests demonstrated robust performance across various metrics, showcasing its adaptability to complex datasets. Logistic Regression, while computationally efficient, displayed some limitations in capturing intricate patterns within the data.

In the discussion, considerations such as model interpretability, computational efficiency, and scalability were examined. The trade-offs between model complexity and performances were deliberated, emphasizing the importance of selecting models aligned with specific healthcare contexts and data characteristics. Challenges encountered during the analysis included addressing imbalances in disease prevalence and optimizing hyper-parameters for each model.

V. CONCLUSION

The aim of this project is to predict disease based on symptoms. The project is set up in such a way that the device takes the user's symptoms as input and generates an output, which is disease prediction. A prediction accuracy probability of 95% is obtained on average. The grails system was used to successfully incorporate Disease Predictor. The problems faced by the medical industry with the unaffordability of the patients to seek dictators and the unavailability of the medical staff can be diminished. This can happen by automating the channelization of the patients to a specialist instead of a generalist. This can happen via the use of a disease prediction system. This system will input the patient's symptoms and produce possible disease as an output with 97% accuracy as compare to earlier models.

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), e0154515. 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. 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", IRACSTInternational 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 multi-linear regression", International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, (2020), 2347-5552. https://doi.org/10.21276/ijircst.2020.8.4.15
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. 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. 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.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303276 |

13. Pisner, D.A. and Schnyer, D.M., Support vector machine, in Machine learning. 2020, Elsevier.101-121.
14. Chen, J., Yu, J., Wen, J., Zhang, C., Yin, Z.e., Wu, J. and Yao, S., "Pre-evacuation time estimation based emergency evacuation simulation in urban residential communities", International Journal of environmental Research and Public Health, Vol. 16, No. 23, (2019), 4599. doi: 10.3390/ijerph16234599.
15. Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., "Disease prediction from various symptoms using machine learning", Available at SSRN 3661426, (2020). https://dx.doi.org/10.2139/ssrn.3661426
16. Taunk, K., De, S., Verma, S. and Swetapadma, A., "A brief review of nearest neighbor algorithm for learning and classification", in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE. (2019), 1255-1260.
17. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S. and Karve, A., "Disease prediction using machine learning", International Research Journal of Engineering and Technology (IRJET), Vol. 6, (2019), 831-833. doi: 10.1126/science.1065467.
18. Ibrahim, I. and Abdulazeez, A., "The role of machine learning algorithms for diagnosing diseases", Journal of Applied Science and Technology Trends, Vol. 2, No. 01, (2021), 10-19. doi: 10.38094/jastt20179.
19. Chhogyal, K. and Nayak, A., "An empirical study of a simple naive bayes classifier based on ranking functions", in AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29, Springer., (2016), 324-331.
20. Kumar, A., Bharti, R., Gupta, D. and Saha, A.K., "Improvement in boosting method by using rustboost technique for class imbalanced data", in Recent Developments in Machine Learning and Data Analytics: IC3 2018, Springer., (2019), 51-66

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details