A Machine Learning Framework for Preliminary Disease Diagnosis from Patient Reported Symptoms

Sanjana N N[1], Sowmya[2], Kavana H C[3], Shravya[4]
[1,3,4]*Student, Dept of Information Science and Engineering, Srinivas Institute of Technology, Mangalore, Karnataka, India.*
[2]*Assistant Professor, Dept of Information Science and Engineering, Srinivas Institute of Technology, Mangalore, Karnataka, India.*
doi.org/10.64643/IJIRTV12I12-206837-459

***Abstract*—Early diagnosis plays a pivotal role in providing proper treatment to the patient. However, many individuals fail to consult a physician due to reasons like unavailability of hospitals near them, lack of sufficient time or high expenses incurred by treatments. In this research paper, an approach is presented to make an early prediction of diseases by analyzing the symptoms input by patients using machine learning techniques. The suggested framework is based on supervised learning algorithms which include Random Forest Classifier, Decision Tree Classifier and Logistic Regression. A list of diseases is provided which includes Malaria, Dengue, Typhoid, Viral Fever, Common Cold, Urinary Tract Infection (UTI). Symptoms entered by user are then transformed into numerical values followed by disease prediction. In addition to the web interface designed in Flask, it provides preventive steps for better health to the patient. Experimental results have shown high performance of the model.**

***Index Terms*—Machine Learning, Disease Diagnosis, Random Forest, Healthcare, Symptom Prediction, Flask**

I. INTRODUCTION

The appearance of innovations within healthcare services has led to substantial changes; however early diagnosis continues to represent an unresolved problem. In their day-to-day life, many people tend to overlook some symptoms, which might be the signs of possible future health issues either because they lack information or are reluctant to see a doctor. Furthermore, many people prefer self-diagnosis and look for potential solutions in the Internet. It means that patients' illnesses get worse over time, and potentially curable diseases might turn into chronic conditions. Hence a system capable of estimating possible health risks according to symptoms is needed. Fortunately, this goal can be accomplished with the help of machine learning technologies. Analyzing data computer systems manage to establish connections between particular symptoms and certain illnesses. As a result, the patient enters his or her symptoms into a machine learning model which estimates the probability of having particular illnesses according to these symptoms. Obviously, this process cannot be called diagnosis, as it does not imply any specific conclusions. Still, the information retrieved could be useful for those people who do not know whether they should see a doctor or not.

For this purpose, an estimator application for disease estimation from symptoms is created in this paper with machine learning techniques. The estimator was developed by training the algorithm through a database with the name of the disease along with symptoms in models like random forest, decision tree and logistic regression. Along with the algorithm, an intuitive user interface was also created to make it easier for the user to interact with the program.

II. LITERATURE REVIEW

During the past few years there have been several works dedicated to the application of machine learning for disease prediction based on the information provided by users about their symptoms. As demonstrated in the previous work of Keniya et al. [1], the use of several algorithms simultaneously is helpful for achieving better prediction quality since different models capture different features. In this way, it is possible to get improved results in some cases, which is why applying only one algorithm may not work

well. Another research carried out by Hakeel et al. [2] considered the performance of various classification models and noticed that more sophisticated ones such as Random Forest yield relatively stable results as compared to other solutions. Thus, the internal model structure is critical for working with medical data in practice.

According to another study conducted by Thirumal et al. [3], it is essential to find efficient ways to manage bigger volumes of data in order to build reliable predictions. Dinesh Kumar and Arthi [4] analyzed various machine learning methods applied in the field of healthcare and found that input data quality is crucial for successful predictions.

Likewise, in their research, Bhanuteja et al. [5] have shown that designing systems which are capable of predicting multiple diseases with the help of the same system is also possible.

With their review Kourou et al. [6] provided insights into the field of medical prediction through machine learning and its possible role in helping with diagnosis at an early stage. Although various researchers have shown that machine learning may be useful in this field, most of these works have been limited to enhancing the accuracy level. User-friendliness has not received much emphasis in this context. Keeping this aspect in mind, the current research attempts to design a user-friendly model.

III. METHODOLOGY

The development of the proposed system proceeds through a series of steps where raw symptom data is translated into meaningful predictions of diseases. Initially a dataset comprising various diseases and their associated symptoms is collected from credible sources. Each entry in the dataset refers to a particular disease with a number of symptoms associated with it in patients. However, before using the data for training, it is cleaned to remove inconsistencies handle any missing data and standardize the data format. Then the symptoms are converted to a structured numeric format using binary coding. In the next step, the data is partitioned into two parts, with the first part used for training and the second for assessing the predictive performance of the model. By partitioning the data in this manner we are able to evaluate the generalization ability of the trained models. Several machine learning algorithms including the random forest, decision tree and logistic regression, are trained on the dataset. These models discover patterns and relationships between the input symptoms and diseases. As a user interacts with the system, they enter their symptoms via a user-friendly interface. The input is fed into the machine learning algorithm in the same form as the training dataset and evaluated using the trained model. Through the established patterns, the model will generate predictions of the most probable disease and display the output to the user. The whole system is designed to be quick and intuitive to ensure users get instant results. In summary this method incorporates data preparation, training, and real time prediction to develop an application that could support preliminary health screening.

IV. SYSTEM ARCHITECTURE

The proposed architecture is split into three main blocks: the interface, the processing block, and the prediction block. The user interfaces with the system through the web-based interface provided, whereby they enter their symptoms. Subsequently, the data obtained is moved to the backend, where it is structured such that it matches the dataset that was used in the training process. Each symptom is classified under a certain group of attributes that will aid in interpreting the input data. Afterward, the processed data is fed into the machine learning algorithms, which are responsible for analyzing the input data and predicting the probable disease. Upon receiving the output data, it is moved back to the interface for display to the user.

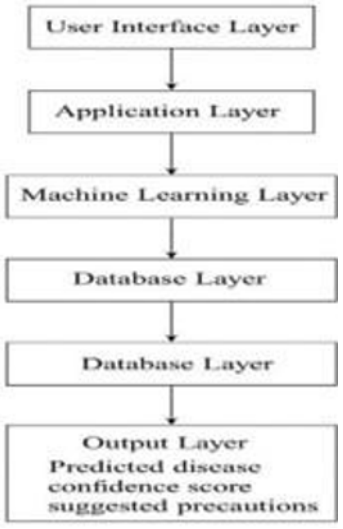

Fig 1: system architecture of disease prediction

V. MODEL TRAINING

The training phase is started after the data preparation phase. The data is divided into two sets such that the majority of data is used for training the models while the remaining data is used for testing. It is useful to assess the performance of the model on unseen data. Training is done for several machine learning models such as Random Forest, Decision Tree, and Logistic Regression using the same data set. In the training phase, the model learns the patterns present in the data that link symptoms to diseases. The model learns to make predictions from the data without memorizing it during the training process. The models are tested using the test data set. The performance of all models is evaluated based on their accuracy. The model that performs better is chosen for deployment in the system.

VI. RESULT AND DISCUSSION

The developed system has been assessed using various symptom inputs, which were intended to measure the ability of the system to accurately predict disease. The data set has been segmented into training and testing sets, and several machine learning algorithms have been tested. The assessment was majorly based on prediction accuracy, consistency, and speed of the algorithm. From the analysis carried out, it was realized that the best results were yielded by Random Forest, which was more accurate than the others. It could manage various symptom input without any difficulties. The Decision Tree algorithm also performed reasonably well except that its results would occasionally be affected by slight changes to the symptoms input. On the other hand, Logistic Regression gave comparatively lower results, especially when the input symptoms were related in some way. Usability has also been tested in this case; with a simple and friendly web-based interface, patients can easily use the system and get instant results within seconds. The prediction output was presented in a clear and understandable manner which increased its user-friendliness and ensured that the system is easy to use for any person. Nonetheless there were several drawbacks during testing. The accuracy of the system greatly relies on the training data set. If there were diseases and symptoms that were absent in the training data set, the predictions would fail to be entirely accurate. Moreover, the system considers the symptoms provided as the main factor of disease prediction and ignores medical history as well as other factors that could impact the final result. The analysis indicates that the proposed method can be a useful tool in predicting certain diseases. It provides an acceptable level of accuracy as well as convenience which allows for the usage of the system in early analysis.

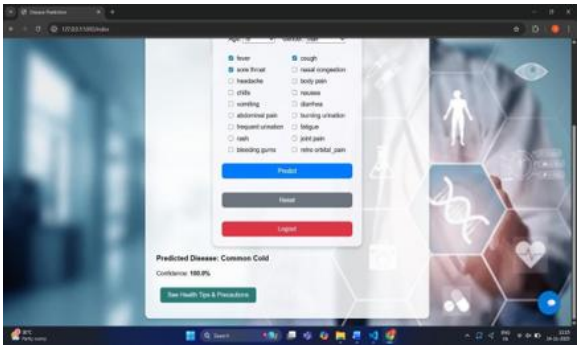

Fig 2: Disease prediction output

VII. USER STUDY

In order to see how efficient the system works in practice; a small user testing exercise was undertaken involving a group of users. The system was handed over to people who knew how to work with a computer and the experiment involved them typing different symptoms and seeing how their results were generated. The main purpose of the test was to see if the system was easy to use and the predictions useful to the user. The majority of the users did not find it hard to operate the interface in addition, the process of getting the results was rather quick, making it ideal for performing quick health checks. It was also noted that the prediction results made sense to the user, especially because they were accompanied by some basic recommendations. Some participants pointed out that more types of symptoms could be included in the system, as well as the recommendations made more detailed. However overall, it should be noted that this system may help the user get a general idea of their condition and decide whether they need to see a doctor or not.

VIII. CONCLUSION

In this paper, the use of machine learning techniques for disease prediction from symptoms submitted by the user is proposed. The system integrates basic data

preprocessing with the use of popular methods resulting in instant analysis and conclusions. With the help of the web interface, the system has been made accessible to users enabling them to utilize the service effortlessly. It can be observed that the results show how effective the proposed system can be for the initial detection of potential health issues. Not only does the model help understand the symptoms but it also urges the users to visit a physician. However, one should remember that the results of the analysis are intended for the preliminary diagnosis and cannot replace professional consultations. In the future, it would be appropriate to expand the number of diseases analyzed, use additional data, and improve the quality of information obtained by the model. In conclusion, it can be said that the proposed system provides a good illustration of the machine learning implementation for healthcare purposes.

ACKNOWLEDGMENT

The authors would like to express gratitude to the entire faculty of the department for their invariable assistance and cooperation throughout the execution of this work. Special thanks are due to the project guide who assisted them with valuable suggestions at every step. The authors would also like to thank their friends for the helpful criticism they provided during this task. We are grateful to all those people who contributed in one way or another to the completion of our project.

REFERENCES

[1] R. Keniya, N. Mehendale, and S. Borate, "Disease Prediction from Symptoms Using Machine Learning," *International Journal of Computer Applications*, vol. 178, no. 1, pp. 1–5, 2020.

[2] M. Hakeel, A. Basha, and S. Khan, "A Study on Symptom-Based Disease Prediction Using Machine Learning Techniques," *International Journal of Advanced Research in Computer Science*, vol. 12, no. 3, pp. 45–50, 2021.

[3] S. Thirumal, K. Nagarajan, and R. Prakash, "An Approach for Disease Prediction Using Big Data Analytics," *Procedia Computer Science*, vol. 50, pp. 101–108, 2019.

[4] D. Dinesh Kumar and V. Arthi, "Prediction of Diseases Using Machine Learning Algorithms: A Survey," *International Journal of Engineering Research & Technology (IJERT)*, vol. 8, no. 10, pp. 200–205, 2019.

[5] T. Bhanuteja, G. Pavan, and K. Reddy, "A Novel Approach for Multiple Disease Prediction Using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 10, no. 2, pp. 150–155, 2021.

[6] K. Kourou, T. Exarchos, K. Exarchos, M. Karamouzis, and D. Fotiadis, "Machine Learning Applications in Cancer Prognosis and Prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[7] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. New York, NY, USA: Wiley, 2013.

[10] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.