Healthcare prediction system Using IOT

Tamboli owaiz, Shaikh Junaid, Mulla Zakriya, Anil Sawant
Department of Electronics and Telecommunication, Trinity College of Engineering and Research, Pune

Abstract- **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 neighbors (KNN), 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: AI chatbot, healthcare, natural language processing, rural healthcare, appointment booking, IOT enabled.

I. INTRODUCTION

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, 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.

Thus, unlike other models, this model concentrates more accurately on these results. The patient can simply enter the disease experienced by him/her, and then this data will be fed into the model, which in turn, provides the possible disease.

The generalized disease prediction architecture that is currently used as of now is not accurate and inconsiderate of the medical history of the patient. The present general model heavily depends on the presence of symptoms and human interaction [8]. All the other methodologies used the symptoms of the patients in the present scenario. For example, the SVM method intakes the symptoms of the patient that have occurred very recently [10]. The generalized disease prediction architecture consists of this methodology only. These methodologies do not

intake the patient's medical history as input data. Due to this, the other generalized methodologies become less effective and have less human interaction. This also affects the accuracy of the model that is presented in the earlier studies. 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 more accurately on these results. The proposed model has the following major contributions:

1. The proposed model has improved Efficiency and accuracy to predict diseases
2. The proposed model is trained on the modified dataset (assigning the weights to the rare symptoms according to the geographical area)
3. The model is tested on real-life symptoms of patients. The remaining section of this paper is structured as follows: section 2 discussed the earlier work done by the authors. Section 3 focussed on the proposed methodology with various methods used to increase the accuracy of the disease prediction model. In section 4 author discussed a comparative analysis of earlier methods and the proposed model. Section 5 concludes the work and is followed by the future scope in section 6.

II. LITERATURE REVIEW

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.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. 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. 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. 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. The KNN method is also used by Kashvi et al. 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.

The method proposed by Pingale et al.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. Also, Gomathy, and Rohith Naidu 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. The method proposed by Chhogyal and Nayak used Naïve Bayes classifier. They have obtained poor accuracy in disease prediction also they are not using the standard dataset for training.

The method proposed by Kumar et al. used Rustboost Algorithm. RUSBoost is developed to address the issue of class imbalance. 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. 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.

The novelty of the proposed work is that tweaking the Radnom forest model by using hyperparameters, 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:

3. 1. Random Forest Algorithm: The random forest produces decision trees from multiple data using their average for regression and most of the voting for categorization . The research reported by Paul et al used the Random Forest Algorithm as the main algorithm.

The random forest algorithm is used to train the model with the dataset which contains a combination of symptoms and the corresponding diseases .The driving force behind using the random forest algorithm is that it has the capacity to handle data sets with continuous variables, as in regression, and categorical variables, as in classification .It produces superior results with regard to classification problems. The working method of the Random Forest is illustrated in Figure 1.

Step 1: Select arbitrary samples from a given data set or training set.

Step 2: This method will create a decision tree for every training data set.

Step 3: Using the decision tree's average, voting will be done.

Step 4: Lastly, select the predicted outcome that garnered the greatest support as the final prediction outcome.

The Random Forest Algorithm analyses the symptoms and geographical region in the provided database to make judgments about a disease. Then it analyzes the outcome with the labels supplied before

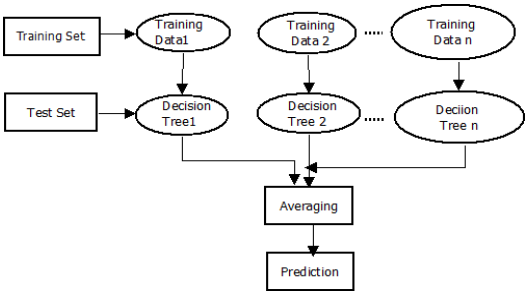

Figure 1. Methodology of Random Forest Algorithm going back to assess the model's reliability. The formula for the random forest algorithm :

$$MSE = 1/N \sum_{i=1}^{N} (f_i - y_i)^2) \quad (1)$$

In Equation (1), N represents the total amount of data points, f_i denotes the model's output, and y_i denotes the real value for data point i. This is used for the calculation of the Mean Squared Error. This method calculates the distance between every node and the expected real value to identify which branch is the best option for your forest. f_i is the decision tree's output and y_i is the value of the data point that you are evaluating at a certain node. You should be aware that while running Random Forests with classification data, you typically use the Gini index, which is the method used to decide the order of nodes on a decision tree branch. Based on the class and likelihood, this method determines the Gini of each branch on a node, showing which branch is more probable. Thus, p_i denotes the class's proportional frequency throughout the dataset, while c is the overall number of classes present.

The architecture of Random Forest Algorithm:

Figure 2 represents the working architecture of the Random Forest Algorithm . As evidently visible, the divided sample of the data is used for further calculation of decision trees at the final which combined serve as a result. The Random Forest algorithm consists of the following steps:

1. Dividing the entire dataset into test and training data
2. Dividing the datasets into multiple datasets
3. Generating Decision trees from each dataset
4. Evaluating these decision trees
5. Concluding the insights generated from the decisions trees
6. Generating the result as an output

3. 1. 1. Advantage of using Random Forest Algorithm

In the database, the author has modified

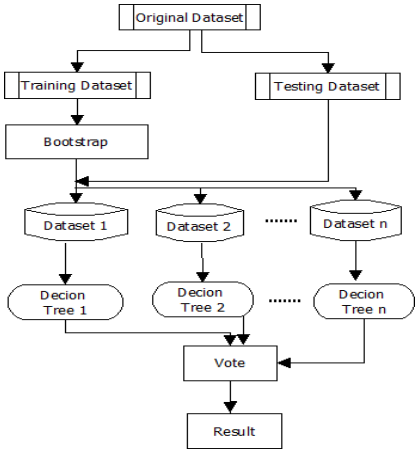

Figure 2. The architecture of the Random Forest Algorithm

the symptoms (inputs) based on the following parameters:

1. Rarity: The rarer a symptom is, the more weight is given to it. Thus, the Random Forest Model predicts a disease more accurately according to the symptoms [1].
2. Location: Some diseases are only bound to happen in a particular geographic location.
3. Thus, the database is set in such a way that the algorithm discards all the diseases that are not present in the inputted location [24].

While training the model, the decision forests that are formed while concluding are pruned as soon as they encounter a weak symptom or a symptom that does not occur in a location. Thus, Random Forest Algorithm minimizes the cost whilst predicting a more realistic model [25].

3. 1. 2. Disadvantage of using Random Forest Algorithm

1. Execution time: It requires huge execution time and space for the compilation of the decision trees [24].
2. Stability: It works better in a stable environment where the dataset is less noisy and subjected to be less dynamic.
3. Overfitting: It may lead to an overfitted model when provided with noise.

3. 2. Long Short-Term Memory Long Short-Term Memory (LSTM) type recurrent neural networks are able to understand order dependency. The LSTM algorithm can be used to calculate and predict disease on the basis of the time-series data of the patient's history of symptoms. LSTM will be used for inculcating the new dataset with the involvement of the pre-trained dataset for increasing the accuracy of the model and discovering new possibilities and parameters.

The inclusion of LSTM will make the prediction of the model more accurate and stable. LSTM will be most accurate when provided a time-series data, which could be inculcated in the future. The input gate is described in the first equation, which also provides the new data that will be added to the cell state. The second is the forget gate, which tells the contents to be removed from the cell state. The final one serves as the output gate that is used to activate the LSTM block's final output at timestamp "t.".

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \quad (3)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (4)$$

Equations (2), (3), and (4) are used for the calculation of the values that are of type-time series generally. Each term in the equations represents the following terms:

i_t: depicts the input gate.

f_t: depicts forget gate.

o_t: depicts output gate.

σ: depicts sigmoid function.

The LSTM model is shown in Figure 3. The above model explains the working of the LSTM algorithm.

3. 3. Support Vector Machine (SVM)

After the result of the value from the LSTM model and the Random forest model, the SVM model will be used to predict whether the result is actually correlated or not. For example, if the LSTM model indicates "Hepatitis" and the Rainforest model also indicates "Hepatitis", we will check with SVM if the results of them are correlated and if it happens due to causation .

In short, SVM will be used to predict the outcome and categorization of the provided inputs depending on the parameters supplied. As a primary approach, the SVM is used in the research publications by Vijayarani, and Dhayanand and Le et al. to predict the outcome using symptoms as input. However, the SVM algorithm

used in our research is solely used for predicting the result between the two parameters. SVM is chosen as the model for the final prediction due to its ability to classify the dataset.

3. 4. Data Transformation About the dataset

The dataset is imported from Kaggle[1]. The dataset consists of 4500+ patients with the parameters as follows: Symptoms (133 columns), Disease (1

column), and Location (1 column).

3. 4. 1. Transformation Methodology

This raw dataset from the Kaggle is then further processed and transformed into numerical values, according to the severity and the rarity of the symptoms. The dataset has been split in proportion for training and testing, 70% of the data consumed for training and 30% for testing, in a ratio of 70:30. The dataset can be further increased with the induction of new patients and new symptoms.

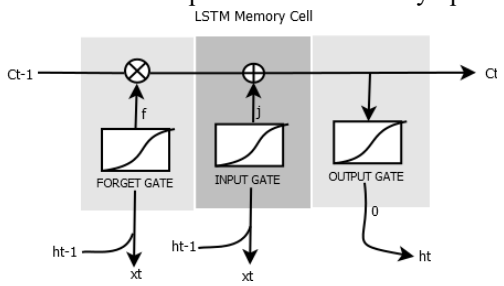

Figure 3. LSTM model [27]

Additional to this data, the model would also required the dataset of the history of the patients. This data would be utilized for training another model for tracking the history of the disease that is and can be suffered by the patient. This dataset would then be trained with Random Forest for concluding. The combination of both these models would help in predicting the disease suffered by the patient. This patient history dataset is not required for prediction since without it, the model would operate on the obligatory model, which uses the disease's symptoms to detect it. As mentioned earlier, the various models have been tested on the modified dataset, finding the methodologies more efficient and accurate.

IV. COMPARATIVE ANALYSIS

To get a glimpse of the difference between the models used by other research papers, Table 1 describes a comparative analysis of earlier methods and the proposed model.

Table 1 explains the comparative analysis of several state-of-the-art methods that are based on the derivation of the disease prediction of a patient using symptoms as input data. The first column represents the reference number, in other words, the serial number of the paper. The second column represents the methodology behind the derivation of the conclusion of the research paper. The basic methods used by the researchers are shown in this column. The research papers listed in the references and in the table have reached conclusions regarding the diagnosis of the disease based on input from symptoms. The third column represents the advantages of using the methodology mentioned in the second column. The advantages are determined on the basis of the analysis of the research paper. Some of these advantages are also unique factors in the research paper and are the factors

that differentiate them from other research papers. The fourth column in the table of the comparative analysis represents the disadvantages of the proposed research papers. These are the limitations that the research papers are not able to solve. However, By solving these limitations, It is analyzed that the proposed model has increased accuracy as compared to earlier state-of-the-art-methods. The fifth column represents the accuracy of the proposed methodology in the research papers. According to the comparison, the initial research paper's highest accuracy was close to 95% which is less than the modified proposed model. The Confusion Matrix for the Random Forest model of the proposed model is illustrated in Figure 4.

Figure 5 shows the comparative analysis of the accuracy of the training models. From the earlier necessities, Naive Bayes Algorithms were best with a model accuracy of 94.8%. Following the Naive Bayes model is a weighted KNN model with an accuracy of 93.5%. The research papers using the SVM model wheres also very close. However, the suggested model, that is Random forest model, yields the most accurate result, 97% as compared to earlier methods.

V. CONCLUSION

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. The proposed model can assist the healthcare industry by:

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%

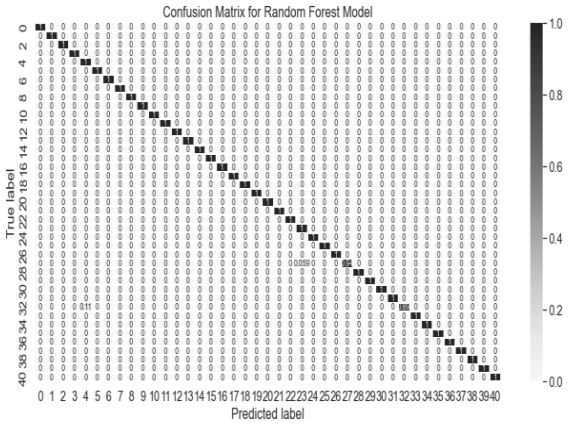

Figure 4. Confusion Matrix of the proposed model

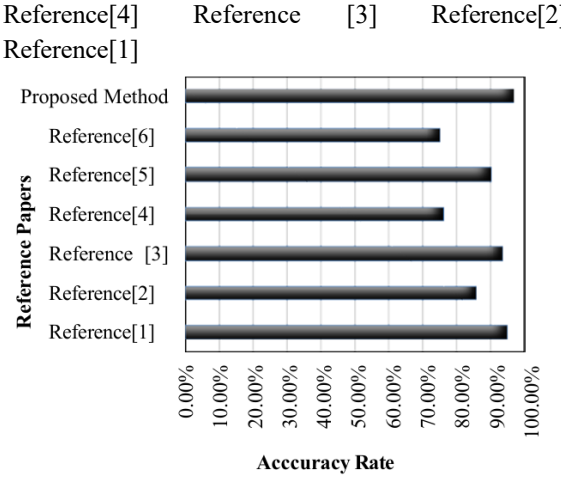

Figure 5. Accuracy of different models

1. Reduction in healthcare costs: By improving patient outcomes and reducing the need for unnecessary tests and treatments, disease prediction applications can help reduce healthcare costs and improve the overall efficiency of the healthcare system.
2. Improved patient outcomes: By providing healthcare providers with valuable insights into a patient's disease risk, disease prediction applications can help improve patient outcomes by allowing for earlier and more effective interventions.
3. Early diagnosis: By analyzing patient data and identifying risk factors for specific diseases, disease prediction applications can help healthcare providers make an early diagnosis, which is critical for improving patient outcomes. At last, we conclude that our model can provide increased accuracy and a reliable model for the prediction of the disease through symptoms.

VI.FUTURE SCOPE

In the future, the model can be used in various sectors and can enhance efficiency by considering more symptoms to predict disease. The model can be used for providing an enhanced, more accurate framework that would lead to a better human disease prediction model.

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", *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. 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.