Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Similar Symptoms Prediction and Disease Prediction Using Machine Learning Techniques

Savita K. Shetty[1] , Shashidhara H S[1]

[1] Dept of ISE, Ramaiah Institute of Technology, (affiliated to Visvesvaraya Technological University, Belagavi), Bengaluru, India
Corresponding author: savita_ks1@msrit.edu

Abstract

Due to the increase in the amount of data growth in the medical and healthcare field, data mining and machine learning technologies can play an important role in the prediction of diseases based on the existing data. Because the average waiting time for patients to meet with their doctor is about twenty to thirty minutes due to many reasons like the doctor may be running late or the patient is taking longer than the time allotted to him, etc. This can be a cause of frustration to the patients as they will have to wait for a considerable amount of time to meet their doctor. This problem can be solved by the proposed model. It makes use of the existing data in an attempt to solve the mentioned problem by significantly reducing the waiting time of the user as it is an application that can be used from the comfort of the home. The application not only provides easy diagnosis to help guide the user but also predicts the symptoms the user might have. The application has two steps. In the first step, a skip-gram model is used. In this model, the dataset is made into the symptom-disease format from the existing disease-symptom format. The symptoms of the disease are predicted in the first step. In the second step, these symptoms are fed into a model that can then predict the disease the person is suffering from. The symptoms of the diseases are predicted using the Random Forest, Decision tree, and Naive Bayes algorithms.

Keywords: Skip-gram, Disease, Symptoms, Random Forest, decision tree, Naïve Bayes.

1. Introduction

In the modern world, there exist a lot of problems related to health. Due to the day-to-day busy schedule and improper lifestyle, the quality of life for people keeps decreasing. Even if the patients fall ill the average waiting time for patients to meet with their doctor is about twenty to thirty minutes. Having to wait a considerable amount of time to meet their doctor can be a cause of frustration to the patients. The evolution of technology plays a major role in the life of people, and can also be used as a means to solve the above-mentioned problem by proposing an application that can be used to predict the disease the user may have from the comfort of his home and recognize the symptoms of their diseases at an early stage.

The skip-gram model or word2vec representation is a very important concept in Natural Language Processing (NLP). Word2vec is a class of models that represents a word in a large text corpus as a vector in n-dimensional space. This model helps to predict and suggest similar symptoms that the patient might be suffering from.

Decision Tree is a very popular technique in machine learning that is used to derive a strategy to reach a particular goal. They are mainly used to visually and explicitly represent decisions and decision-making. This method can be utilized to predict the disease that the patient is suffering from based on his symptoms. Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Snehal Bhoir et al. [1] gives a detailed comparison of the three-word embedding models: Continuous bag of words, Skip-gram, Glove (Global Vectors for word representation), and HellingerPCA (Principal Component Analysis). It analyses the three models and concludes that GloVe is the best-suited model. This is because GloVe is scalable i.e., it can work with both large and small Corpora. Dhiraj Dahiwade et al. [2] paper context-aware clustering is used. It uses the GloVe dataset to generate word embedding for the words in the data set that needs to be clustered. The data set objects are then clustered using k-means ++. The above-mentioned method is better than any previously used method because it concerns and preserves the dependency within the data at a more refined level than in the case of the k-means algorithm. It is also more efficient as it is faster and more precise than the k-means algorithm. K-means ++ can be used for confidential data as well because it can work on synthetic data as well.

In [3] authors explored the SD Skip-Gram model, for medical concepts. It incorporates the synthetic dependencies into the basic Skip-Gram model. When the vector representations generated through this model are compared to the ones generated by the SD Skip-Gram model, it can be concluded that the SD Skip-Gram model can identify more related diseases and symptoms. They also compared the Skip-Gram and SD Skip-Gram models. They are evaluated on the medical disease and symptom dataset. The results show that the performance of the SD Skip-Gram model's is good in recognizing the relations between common words. This model can recognize the related diseases and symptoms for the selected diseases more closely. In [4], the authors suggest a new method for recognizing diseases and predicting the time it takes to cure them by considering the symptoms of the diseases. In this method, different coefficients are assigned to different symptoms. They also assign a severity score to each symptom. The basis of this method is that every disease symptom has a unique effect on the time taken for recovery. This system then exploits the patterns for predicting the time taken to be cured and the associated diseases. The symptoms and diseases are associates with each other by matching the symptoms with entries in the database, then the common symptoms and the possible diseases are selected for further processing. But this system is also on the previous users as it should be able to distinguish between the real and the fake ones.
In [5] authors implemented GloVe with RNN. Foer every word GloVe generates 50 dimensions' word vector. It is then fed to the Recurrent neural network layer. The RNN has only outputs i.e., positive and negative (binary classification). This model can be trained and used to perform the sentiment analysis. The disadvantages of this method are that the similarity is difficult to explain as there is not much evidence and that is implemented only for binary classification. In [6] authors described a process in which the evidence that illustrates the similarity of two terms can be drawn out automatically. The idea of "similarity" by the combination of "commonality" and "aligned difference" is determined based on a human's comparison process. They also list several criteria that guarantee the quality of the evidence such as relevance, semantic similarity, relational similarity, and systematicity. This process returns the fully unsupervised set of evidence that is vital to judging the similar characteristics or features of any two given terms. But authors do not test the methods in other situations, like whenever using different corpus and with different word embedding methods.
In [7] authors provide a comprehensive mathematical analysis of the skip-gram model. They derive various formulas that are connected to the SGNS model. Then the best solution on the vectors is provided and is supported by the results from the experiments performed. The authors do not provide a deeper analysis that shows why analog relationships are satisfied by the word2vec vectors. It also doesn't come up with a composite language model applying expression embedding, which can capture the meanings of expressions and semantic units.
In [8] authors focused on the vectorization technique They built a framework for the word segmentation using a medical dictionary and a Markov model. Then the Skip-Gram model learns each text word into a 128-dimension vector. By doing this vectorization on the Skip-Gram model can improve the performance of the CNN classifier and the accuracy is also increased. Authors can implement only for Chinese data but can be used as a model to implement other languages as well. In [9], a new weighting method is described for the ensemble prediction in which the weighted average of outputs of multiple models is used. Models with relatively high prediction accuracy are given large weights and fluctuation of prediction errors is suppressed and high prediction accuracy is expected. More advanced operator assistance is promoted with the anomaly symptom detection technology based on the proposed ensemble prediction. This method is to determine weights in proportion to the reciprocal of RMSE (Root mean squared error) was adopted because in prediction, a model with high prediction accuracy i.e. the small error is largely weighted and those with low accuracy i.e. large error are less weighted. Therefore, it is considered that weights are determined by the error of each prediction model. Then, because weights are in general normalized as a sum of weights to be one, just relational magnitude of error is necessary to determine weights.
In [10], the authors are mainly focusing on the prediction of heart disease. The Random Forest algorithm is used. The chi-square technique is used for feature selection[14]. This combined approach attained an accuracy of 83.70%. It also provides comparisons between different classification techniques. It also considers various parameters for heart disease data set such as Sensitivity. Specificity, Disease prevalence etc. Authors concentrated only on one disease i.e., heart disease. In [11][13], the authors implemented based on the medical diagnosis history dataset of individual person and then predicting the disease risk. Here a total of eight disease categories can be predicted. Overall, the Random Forest method outperformed on different classification algorithms. By combining repeated random sub-sampling with Random Forest, the class imbalance problem could be overcome and was able to achieve promising results. The result was a prediction of eight disease categories with an average AUC of 88.79%. Authors tried to predict more than one disease.

2. Materials and Methods

2.1 Proposed Architecture

The proposed architecture in Figure 1, divides into two main modules. The two main modules include Symptom-Disease lookup System and Disease prediction System. There is a view component involved in this architecture.

The view component of the architecture involves the User Interface components of the model. The two other modules receive the incoming inputs, request from the view component, and manipulate the data resources through building the models and interact with the views to render the final output.
The view component of the architecture involves a User Interface to users to interact with the system. The view component includes various fields to get inputs from the user. The view component has the login page where the user can enter the username and password. There is a particular flow for the execution of two modules. There are two fields to perform the task of Symptom prediction and Disease prediction. After logging in, the user can enter the symptom. When the user submits the request by clicking on the predict similar symptom button the symptoms similar to the input is given by the user are listed as a list. The user has to select symptoms that he/she is showing. If the symptoms are not in the list, then the user has the provision to enter the symptoms in the text box given. Sometimes the users are not familiar with the medical terms. So, the user can drag and drop the symptoms that he is unaware of. The output received from the first module is input to the second module. The second module involves performing sub-tasks. There are three separate buttons for these sub-tasks. Based on the sub-task selected by the user corresponding result is visualized on the view component.
The two main modules of the proposed architecture use machine learning tasks. In the first module, the dataset which is obtained from an online repository is pre-processed to bring it into a particular format. The dataset initially obtained is in text format. The main objective of this module is to find similar symptoms to the input symptom. Based on the GloVe representation of the words the similar symptoms are obtained. In the second module, the same dataset is pre-processed to a categorical format. Different classification models are used to predict diseases. This prediction system predicts multiple diseases. The classification model implemented here is Decision Tree, Naïve Bayes, and Random Forest.

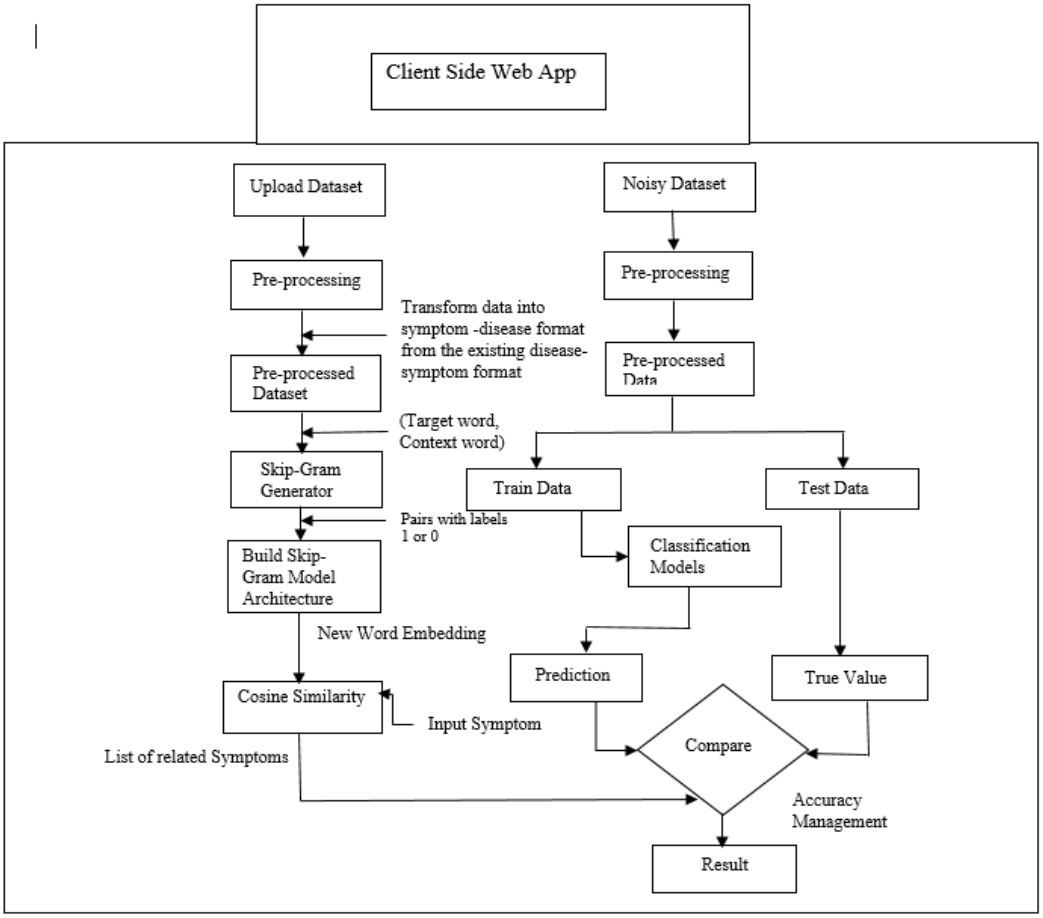

Figure 1. Proposed Architecture

2.2 Construction of Symptoms Prediction Model

2.2.1 Data Pre-processing:

The first step is to do text data pre-processing which involves the deletion of special characters, additional whitespaces, and digits. Additionally, the text corpus is converted into lower case.

The dataset consists of data in the form of Disease-Symptom format that is for a particular disease, a list of symptoms about the disease has been specified. As the model tends to predict a list of similar symptoms based on the symptom mentioned by the user, the concept of target word and context word has been used to train the model. Here the symptoms are considered as the target words whereas the diseases are the context words that will be feed into the skip-gram model, which predicts the target words based on the context words that is the similar symptoms are predicted with the help of the disease that they are associated with.

2.2.2 GloVe for word representation:

GloVe can be used to represent the words in the vector form. Which also stands for global vectors. These global vectors are obtained by training global word-word co-occurrence statistics from a corpus, which results in linear substructures of the word vector space. The pre-trained 50-dimensional GloVe word vector has been used. The representations are stored in a dictionary, where the words are the key and the corresponding vectors are the values, which is used to train word embeddings on top of existing GloVe representation with the use of skip-gram model

2.2.3 The Skip-gram model:

The Skip-gram model is used to train word embeddings on top of the existing GloVe representation. Each symptom has a disease associated with it, which is used as (target word, context word) pair for a skip-gram generation. The most related words for a given word can be found with the help of an unsupervised learning technique called Skip-gram. This model predicts similar symptoms i.e., the context words from a target word, which is the symptom specified by the user.

The model has to be trained to identify what is contextual and what is not, for which the model is feed with input [(target, context), 1] pairs, which are positive samples. The negative samples, [(target, non-context),0] are also feed into the model. The negative label 0 shows that it is a contextually unrelated pair.

Regarding the dataset, the positive input samples are [(symptom, disease associated with the symptom), 1] and the negative input samples are [(symptom, disease that is not associated with the symptom), 0]

2.2.4 Building the skip-gram architecture:

To begin with, the corpus vocabulary is built, by assigning a unique identifier to the unique words from the vocabulary. The required skip-gram pairs have been generated that is the pair of target and context word with label 1 and the pair of target and non-context word with label 0.

The skip-gram model architecture is built on Keras on top of Tensor Flow. Each of the target and context or non-context word pairs are passed to:

i) ***Embedding layer***: This layer is initialized with the word embeddings from the GloVe vectors. Once the word embeddings for the target and the context word is generated, it is passed to the Merger layer.

ii). ***Merger layer***: Here, the dot product is computed for the two vectors, i.e., the vectors for the target and the context word generated in the embedding layer. The dot product is passed to a dense sigmoid layer.

iii). ***Sigmoid layer***: Depending on if the pair of words are contextually relevant or just random words, this layer predicts either 0 or 1. Then the output is matched with the actual relevance label that was generated earlier, the loss is computed and the backpropagation is performed with each epoch to update the embedding layer in the process. The model is now trained with skip-grams. The model is now run on the complete corpus for 25 epochs. When the model is trained, similar words will have similar weights based on the embedding layer.

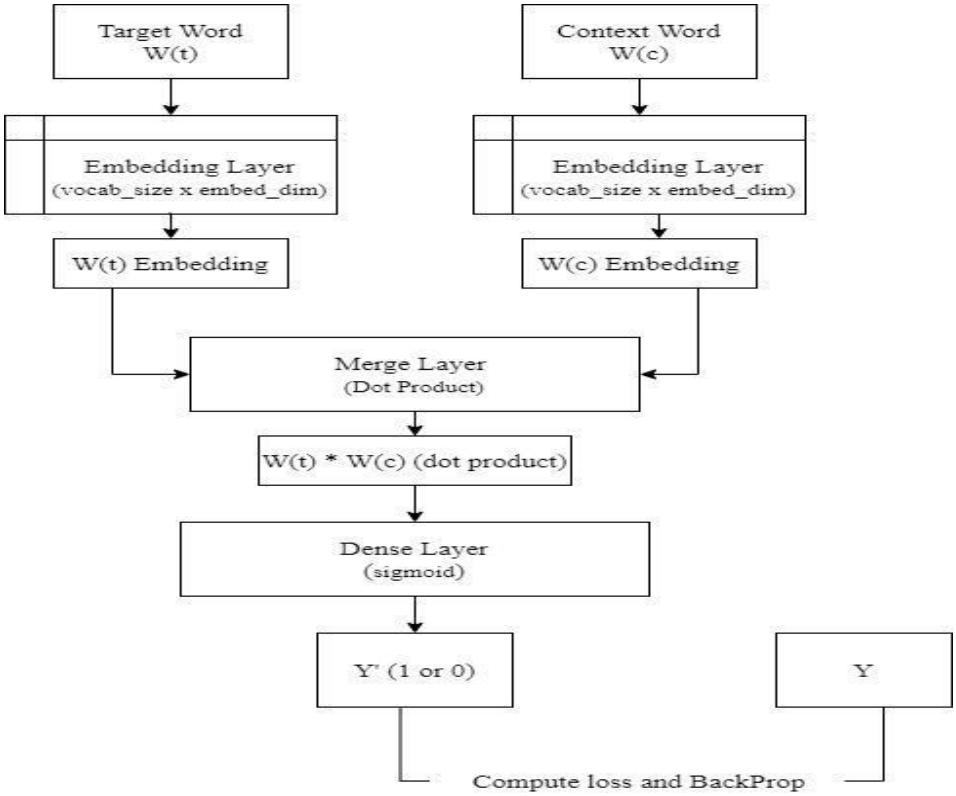

Figure 2 Skip-Gram Deep Learning Model

2.3 Prediction the similar symptoms

The new word embeddings are obtained on top of existing GloVe vectors with the help of a skip-gram model. Once the user enters a symptom, it is looped through the set of all symptoms in the dataset to find out the cosine similarity between the embeddings of the given symptom and the current symptom in the loop, and then the symptoms with high similarity score are generated as output.

2.4 Construction of Disease Prediction Model

Once the user selects the symptoms from the list of similar symptoms generated from the skip-gram model. The selected symptoms are fed into a supervised machine learning classifier model.

The dataset to train the classifier consists of the symptoms as the attributes and the values are either 1 or 0, the value is 1 if the symptom is associated with the particular symptom otherwise the value is 0

2.4.1 Disease Classification

Three supervised algorithms that are decision tree classifiers, random forest, naive Bayes have been used. The three models are compared based on the accuracy obtained and the one with the highest accuracy is used to predict the disease. In this case, the decision tree classifier produces the most accurate results.

2.5 Prediction of the disease

The dataset is split into training and testing data. The model is trained upon the training dataset and tested based on the data available in the testing data. Once the model is trained, the input symptoms from the user are fed to the trained model which predicts the disease the user is suffering from, which is the result of the proposed model.

3. Results and Discussion

3.1 Similar Symptoms Prediction

The dataset consists of data in the form of Disease-Symptom format that is for a particular disease, a list of symptoms about the disease has been specified. The dataset consists of 148 unique diseases and 405 unique symptoms. The new word embeddings are obtained on top of existing GloVe vectors with the help of a skip-gram

model. After running the model for about 25 epochs the model generated good results. The cosine similarity between the user entered symptom and symptoms that are looped through the dataset is calculated. Symptoms having cosine similarity greater than the similarity score (i.e. 0.6) are found. Refer to for similar symptoms prediction.

3.2 Disease Prediction

The dataset used for training the models consists of 2113 samples of 148 unique diseases. The same dataset is used in both modules. There is a difference in the way the dataset is preprocessed. This is a basic classification problem. The classifier models are built using Multinomial Naive Bayes, Decision Tree, and Random Forest. Each of the models was trained and tested individually on the dataset. Accuracy is considered as the evaluation criteria for performance evaluation.

Model Names	**Accuracy**
Multinomial NB	75.28%
Decision Tree	78.35%
Random Forest	82.02%

Table 1. Accuracy of Different Models

The Random Forest model outperformed all the other classifier models by obtaining an accuracy of 82.02% while the other models achieved 75.28% and 78.35% respectively.

Apart from these accuracy measures, the models are evaluated for performance in terms of macro-Precision, recall, and F1 scores. Macro F1-scores average the F1-scores for individual Disease classes.

micro avg	0.753	0.753	0.753
macro avg	0.679	0.695	0.683
weighted avg	0.738	0.753	0.742

Figure 3. Average of precision, recall, and F1 score for Multinomial NB

micro avg	0.784	0.784	0.784
macro avg	0.721	0.759	0.731
weighted avg	0.748	0.784	0.757

Figure 4. Average of precision, recall, and F1 score for Decision Tree

micro avg	0.809	0.809	0.809
macro avg	0.743	0.769	0.749
weighted avg	0.786	0.809	0.791

Figure 5. Average of precision, recall, and F1 score for Random Forest

3.3 Comparison of Models

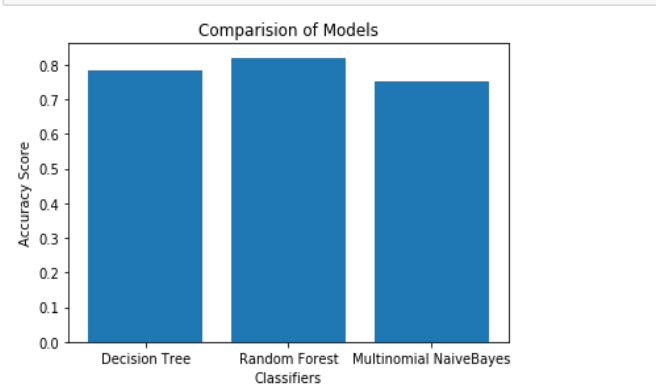

Figure 3. Accuracy of three models

4. Conclusion and Future Work

The models that have been developed are to successfully predict diseases as well as their symptoms. To predict symptoms, GloVe and Skip Gram techniques have proven to be very effective as it gives us a list of all similar words(symptoms) related to a particular symptom. For disease prediction, three models have been evaluated for their accuracy to be used in the project namely - Naive Bayes, Decision Tree, and Random Forest. The machine learning technique ultimately used is Random Forest due to its higher percentage of accuracy in comparison to the other two models. The accuracy percentage achieved is 82.02, which speaks for its credibility to be used personally and professionally and establishes the fact that results achieved are up to the standards.

The base and the backdrop provided by this model opens the scope for more advanced disease prediction depending on the specific data. The work done in this project could as well be expanded to cover more sophisticated techniques to collect health data and employ them in the disease prediction model for more accurate prediction. The same concept would also be used for hardware implementation to develop wearables and other devices that help monitor the symptoms of patients.

References

[1] S. Bhoir, T. Ghorpade and V. Mane, "Comparative analysis of different word embedding models," 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2017, pp. 1-4, doi: 10.1109/ICAC3.2017.8318770.

[2] Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1211-1215. IEEE, 2019.

[3] A. Gupta and B. K. Tripathy, "Implementing GloVe for context based k-means++ clustering," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 1041-1046, doi: 10.1109/ISS1.2017.8389339.

[4] Shankar, Mani, Mayank Pahadia, Divyang Srivastava, T. S. Ashwin, and G. Ram Mohana Reddy. "A novel method for disease recognition and cure time prediction based on symptoms." In *2015 Second International Conference on Advances in Computing and Communication Engineering*, pp. 679-682. IEEE, 2015.

[5] A. Gupta and B. K. Tripathy, "Implementing GloVe for context based k-means++ clustering," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 1041-1046, doi: 10.1109/ISS1.2017.8389339.

[6] Zhang, Yating, Adam Jatowt, and Katsumi Tanaka. "Towards understanding word embeddings: Automatically explaining similarity of terms." In *2016 IEEE international conference on big data (big data)*, pp. 823-832. IEEE, 2016.

[7] Zhang, Canlin, Xiuwen Liu, and Daniel Biś. "An analysis on the learning rules of the skip-gram model." In 2019 international joint conference on neural networks (IJCNN), pp. 1-8. IEEE, 2019.

[8] Zhou, Zhiyang, Bo Fu, Hang Qiu, Yanlong Zhang, and Xiaobing Liu. "Modeling medical texts for distributed representations based on Skip-Gram model." In *2017 3rd International Conference on Information Management (ICIM)*, pp. 279-283. IEEE, 2017.

[9] Murakami, K., Suzuki, S., & Matsui, T. (2017, September). Anomaly symptom detection using ensemble prediction based on new weighting method. In *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* (pp. 1144-1146). IEEE.
[10] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. *Journal of network and innovative computing*, *4*(2016), 175-184.
[11] Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*, 1-13.
[12] Balakrishna, Tilakachuri, B. Narendra, Mooray Harika Reddy, and Damarapati Jayasri. "Diagnosis of chronic kidney disease using random forest classification technique." *Helix* 7, no. 1 (2017): 873-877.
[13] K. V, H. K. N, D. G, K. S. A, N. M and K. Gopal, "AIRFACTOR- Bangalore Based Air Pollution Monitoring and Prediction Application Using Machine Learning," 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2023, pp. 1-6, doi: 10.1109/NMITCON58196.2023.10276050.
[14] K. R. Swetha, N. M, A. M. P and M. Y. M, "Prediction of Pneumonia Using Big Data, Deep Learning and Machine Learning Techniques," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1697-1700, doi: 10.1109/ICCES51350.2021.9489188.