IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Disease Prediction from Symptoms Using Machine Learning

[1]Prathamsesh J Lonare, [2]Swapnil S. Kumbhare, [3]Mohak S. Talodhikar, [4]Suyog S. Madavi, [5]Dr.Vanita Buradkar

[1]Student Computer Science and Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur

[2]Student Computer Science and Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur

[3]Student Computer Science and Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur

[4]Student Computer Science and Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur

[5]Assistant. Professor, Department of Computer Science of Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur

ABSTRACT

This Project is all about creating a smart system that can predict diseases like dengue, malaria, and typhoid by analysing symptoms reported by patients. From a predetermined list of symptoms and the associated diagnoses, it employs three distinct machine learning algorithms—Random Forest, Decision Tree, and Support Vector Machine to identify diseases. Each algorithm has gone through extensive training and testing to ensure its accurate and reliable in making predictions. The Random Forest algorithm exhibits superior accuracy in disease prediction when compared to both the Decision Tree and Support Vector Machine models. It reliably produces consistent results by effectively recognizing the appropriate disease based on the symptoms provided by the user, thereby establishing itself as the most efficient model within our framework.

Keywords: Machine Learning, Disease Prediction, SVM, Random Forest, Decision Tree, Classification

1. INTRODUCTION

1.1 *Motivation*

Machine Learning (ML) has emerged as a crucial technology in disease prediction, leveraging vast amounts of historical data to improve healthcare outcomes. This technology enables computers to identify patterns and make predictions based on both structured and unstructured medical data. The creation of an ML model consists of two key phases: training, where the model enhances its predictive capabilities using a specific dataset, and testing, which evaluates its accuracy with new information. In the realm of disease prediction, ML algorithms analyse patient symptoms, thereby aiding healthcare professionals in achieving more precise and efficient diagnoses.[1]

The importance of ML in the healthcare sector is underscored by its ability to quickly and accurately process large volumes of data. As data collection methods advance, healthcare data has become increasingly intricate and plentiful, often including both unstructured text and structured records. By examining this data, ML models can deliver timely insights that are essential for effective decision-making. This functionality enables physicians to make informed diagnoses and tailor treatments to meet the unique needs of individual patients, ultimately enhancing healthcare services.[2]

Our study makes use of several machine learning methods, such as Classifiers SVM, RF, and DT, which are well known for their ability to analyse medical data and make precise disease predictions. The incorporation of ML into healthcare streamlines the analysis of patient data, which not only aids in diagnosis but also encourages proactive monitoring and personalized healthcare approaches. As the healthcare industry increasingly adopts these sophisticated ML techniques, the potential for improved patient care and enhanced healthcare delivery becomes increasingly achievable.

1.2 Aim

This study aims to evaluate the idea which supervised machine learning (ML) algorithms can enhance healthcare by making it possible to diagnose diseases accurately and quickly. This study focuses on examining research that employs multiple supervised ML models for each disease recognition challenge. This methodology ensures greater comprehensiveness and accuracy, as assessing the performance of a single algorithm across different study contexts can introduce bias, leading to inaccurate results. A variety of illnesses will be examined, with a particular emphasis on malaria, typhoid, and dengue. Based on reported symptoms, our design assesses the efficacy of several methods, including DT, RF, and SVM, in predicting diseases including dengue, malaria, and typhoid. This literature review aims to determine the most effective machine learning models tailored to each individual disease.

This project carries significant value in the healthcare sector by leveraging machine learning models for predicting diseases like dengue, malaria, and typhoid. It contributes to the modernization and improvement of healthcare services by offering efficient, accurate, and accessible diagnostic support. All models have the likely to enhance the disorder process and lay the groundwork for advancements in personalized healthcare, enabling tailored treatment strategies based on predictive data. Overall, integrating machine learning into disease prediction marks a pivotal step toward better health outcomes and more efficient healthcare delivery systems.

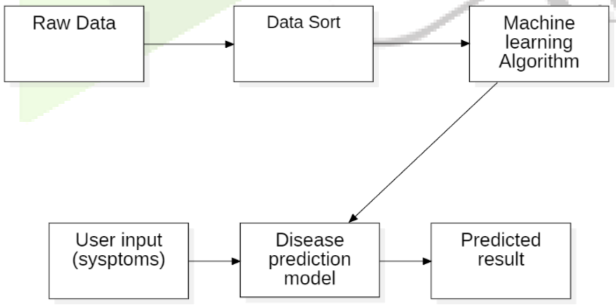

Figure 1. Block Diagram Proposed Architecture

2. LITERATURE REVIEW

In addition to including a module for hospital referrals, this study investigates the use of machine learning approach in creating a system for forecasting diseases such as dengue, malaria, and typhoid. Based on a person's stated symptoms, a variety of classification algorithms, including the Random Forest Classifier, Support Vector Machine, and Decision Tree, are used to forecast the likelihood of their sickness. The most appropriate hospital type for consultation is then suggested. The system is designed for use by end-users, featuring an interactive front-end interface that connects to a server. This innovation has the potential to significantly influence future medical practices. However, the inherent complexity and diversity of diseases may lead to challenges regarding accuracy and potential biases in the data utilized for training the algorithm.[3]

The study "Disease Prediction using Machine Learning" explores how machine learning techniques can be used to predict illnesses based on patient symptoms. It describes how to diagnose diseases like dengue, malaria, and typhoid by using models like Random Forest, Support Vector Machine (SVM), and Decision Trees. It also emphasizes the necessity of accurate sickness prediction in the medical field.[4]

The Decision Tree model leverages the CART algorithm to perform classification, while the Random Forest model applies an ensemble learning approach, combining multiple decision trees to enhance prediction accuracy in disease detection. Furthermore, SVM is applied to classify diseases by maximizing the separation between different classes. The primary aim of this project is to accurately predict diseases based on symptoms provided by users, evaluate the performance of these classification techniques, and analyse how variations in symptom patterns affect the outcomes of disease predictions.[5]

The researchers concluded that the proposed model demonstrated enhanced accuracy, which can be ascribed to its capacity to discern complex nonlinear relationships within the feature space. Furthermore, the Decision Tree (DT) effectively highlights key features, offering a deeper insight into the disease and thereby enabling precise predictions for multifaceted diseases.[6]

This study introduces a framework aimed at the early prediction of diseases through the utilization of an ensemble model that integrates Support Vector Machines (SVM). It demonstrates the efficacy of this methodology across various diseases, which may encompass those of particular interest to you. The paper offers perceptive perspectives on the use of machine learning in the healthcare sector for the early identification of different illnesses, stressing the significance of precise forecasts and prompt actions to enhance patient outcomes.[7]

The conventional approach to diagnosis typically requires a patient to consult a physician, undergo various tests, and ultimately arrive at a conclusion. This procedure may require a significant amount of time. The proposed project introduces an automated system for disease prediction, designed to accelerate the preliminary phase of disease identification that depends on user input. The system is structured so that when users interact with the chatbot, they can opt to receive an estimate or prediction of their potential illness based on the information they provide. Nonetheless, it is crucial to acknowledge that the accuracy of predictions may diminish when a limited number of symptoms are documented. [8]

This research utilizes the Support Vector Machine methodology to assess the accuracy rate. The evaluation of the system's performance is based on metrics such as accuracy, specificity, and sensitivity. The proportion of correctly and incorrectly classified instances is presented as a percentage of the sample data. The models examined in this research include SVM, Random Forest, Decision Tree. A robust multi-process approach that integrates the decision tree algorithm with clustering techniques was implemented in this system. This methodology aids in the development of a predictive system for cardiac arrest. It can be concluded that the Random Forest classifier algorithm demonstrates superior accuracy and efficiency compared to the other models.[9]

3. DATASET DESCRIPTION

The dataset employed in our project, titled "Disease Prediction System Using Random Forest, Decision Tree, and SVM Algorithms," consists of a meticulously organized collection that targets three specific diseases: dengue, malaria, and typhoid, alongside 132 distinct symptoms. Because this meticulously curated dataset accurately depicts real-world healthcare scenarios, our machine learning algorithms can accurately forecast these illnesses from user-provided symptoms. A strong emphasis has been placed on ensuring data quality and integrity to facilitate dependable predictions. To uphold consistency and improve model performance during both the training and testing stages, a range of comprehensive preprocessing techniques has been implemented, including the management of missing data, encoding of symptom information, and normalization of inputs.

4. PROPOSED METHODOLOGY

Our project employs a systematic methodology as follows:

4.1 ***Motivation*****:** A detailed dataset encompassing symptoms and disease classifications, as indicated in the 'prognosis' column, is utilized. This dataset comprises a range of symptoms, including fever, headache, muscle pain, nausea, among others, which are associated with diseases such as Dengue, Malaria, and Typhoid.

4.2 ***Data Preprocessing*****:** The subsequent phase of our project focuses on the cleaning and preprocessing of the dataset. This process entails addressing any gaps in symptom data and maintaining the dataset's consistency and integrity. We will apply preprocessing techniques such as encoding categorical symptoms and standardizing the input features to improve the dataset, ensuring efficient training of the machine learning models for disease prediction.

4.3 ***Feature Selection*****:** Analyze dataset to determine the symptoms associated with the diseases under consideration, specifically Dengue, Malaria, and Typhoid. Employ feature engineering techniques to emphasize those symptoms that are most representative of the three diseases in question.

4.4 *Model Selection*

1. Random Forest Classifier: Employs a collection of decision trees to enhance classification precision and mitigate the risk of overfitting. It is particularly effective for managing intricate datasets characterized by numerous features (symptoms).

2. Support Vector Machine (SVM): Utilizes kernel methods (either linear) to identify the most effective decision boundary separating different classes. This approach is advantageous when the symptoms do not exhibit linear separability.

3. Decision Tree Classifier: Delivers a model that is easily interpretable, allowing for straightforward understanding of predictions based on established decision rules. The system employs the Classification and Regression Tree approach to facilitate the splitting of nodes for classification purposes.

4.5 *Model Training*:

1. **Random Forest:** This technique employs an ensemble learning strategy by generating multiple decision trees and aggregating their predictions through a majority voting mechanism. This approach significantly improves accuracy, minimizes the risk of overfitting, and adeptly captures the correlations between symptoms and diseases.

2. **SVM:** This method utilizes a linear kernel-based classification framework to distinguish between different classes of data. Given that the data is linearly separable, SVM effectively separates the symptom patterns into distinct categories, facilitating accurate disease prediction.

3. **Decision Tree:** This method applies the Classification and Regression Tree algorithm to divide the dataset into subsets by analysing the attributes of the reported symptoms. Although it offers high interpretability, strategies are implemented to address its propensity to overfit when dealing with noisy data.

Prediction and Testing: The trained models are employed to forecast diseases based on symptoms provided by users. To ensure the models' reliability and robustness, they undergo rigorous testing with a dedicated dataset to validate their predictive capabilities.

Evaluation Metrics: The model performance is evaluated using accuracy, to ensure their effectiveness. This detailed assessment helps identify areas for improvement, enhancing the models' reliability and real-world applicability in disease prediction.

Deployment: Following the attainment of consistent performance across all evaluation metrics, the models are incorporated into a user-friendly interface for practical use. Users can input their symptoms, and the system predicts diseases such as dengue, malaria, or typhoid using reliable algorithms Random Forest, Decision Tree, and SVM. This systematic approach ensures that the project is designed to provide accurate and trustworthy disease predictions while remaining accessible and efficient for end users.

5. PROPOSED METHODOLOGY

5.1 Random Forest Algorithm

Random Forest also considered one of the most efficient algorithms for supervised classification. This ensemble learning method exhibits enhanced accuracy compared to the other classification techniques, such as bagging and boosting. Its effectiveness is especially prominent in real-world applications, where it provides reliable solutions for handling missing data and shows a strong resistance to overfitting. The underlying mechanisms of the Random Forest algorithm are elaborated in Algorithm.[10]

The Random Forest model exhibits outstanding performance in our disease prediction system, achieving flawless accuracy on both the training and testing datasets, highlighting its robustness and dependability. By utilizing ensemble learning, Random Forest builds multiple decision trees and generates predictions based on the majority vote of these trees, ensuring accurate outcomes. By employing bootstrapping and random feature selection, it ensures diverse tree generation, effectively reducing the risk of overfitting while enhancing predictive accuracy. The Random Forest algorithm produced extremely accurate predictions in this study by successfully capturing the relationships between the target diseases (typhoid, dengue, and malaria) and the reported symptoms. These flawless ratings highlight its capacity to differentiate between various illnesses using the given symptom information, confirming its appropriateness for reliable use in medical diagnostics.[11]

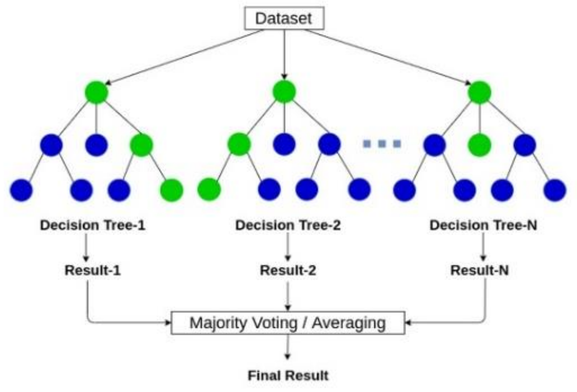

Figure 2. Random Forest Algorithm

5.2. *Decision Tree Algorithm*

Starting with a single node, a decision tree is a hierarchical structure that grows into multiple possible outcomes. In contrast to linear models, decision trees represent a form of supervised learning capable of capturing non-linear relationships. The dataset is partitioned into homogeneous subsets by identifying the most significant split among the input attributes. This splitting criterion is determined through the application of several algorithms, including the Gini Index.[12]

The decision tree algorithm falls within the realm of supervised learning methodologies and is suitable for both regression and classification applications. It utilizes a tree structure to facilitate predictions, commencing with a root node. This initial node is then partitioned according to the most influential input feature, and this iterative process of division persists. The procedure continues until all input data is fully represented, with the terminal nodes indicating the respective weights. To ensure its performance of the model successful implementation in actual healthcare applications, it is critical to identify and correct any biases in dataset and evaluate across various demographics. The accuracy and reliability of the model will be further enhanced by ongoing assessment and updating with fresh data.[13]

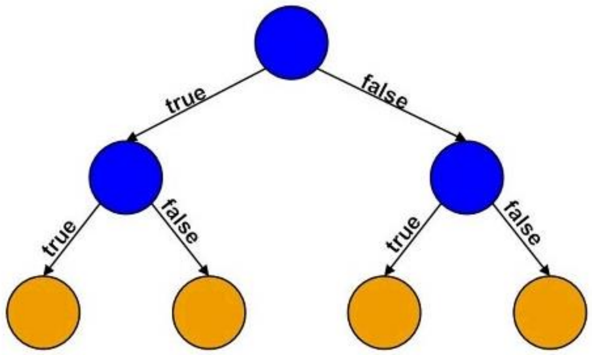

Figure 3. Decision Tree Algorithm

5.3. *Support Vector Machine*

Our disorder prediction system uses Support Vector Machines, which are supervised learning algorithms, both for classification and regression problems. These algorithms categorize data points situated in a multidimensional space by utilizing parallel lines known as hyperplanes. The process of classifying these data points focuses on maximizing the margin between the hyperplane and the nearest data points. Various kernel functions exist for the purpose of mapping both linear and nonlinear data points within a multidimensional space to facilitate their separation. In this analysis, we have exclusively utilized the Linear and Radial Basis Function kernels.[14]

The goal of SVM, a type of supervised machine learning, is to identify the optimal hyperplane for class differentiation. With an accuracy of 94% on both the training and testing datasets, the SVM model fared better in this project than the Decision Tree. This suggests that the model outperformed the decision tree model in terms of generalization and effectively categorized diseases according to their symptoms.[15]

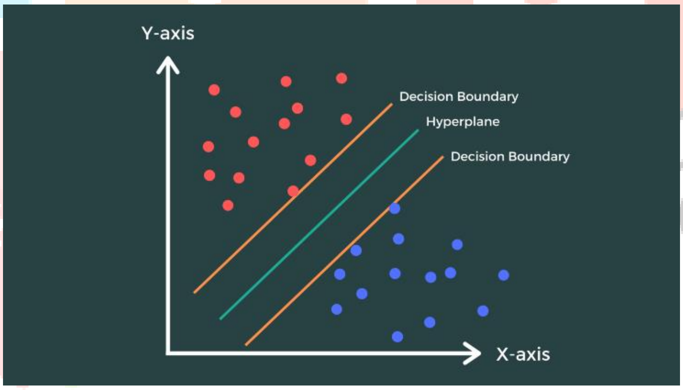

Figure 4. positive vs negative test case

6. RESULT AND EVALUATION

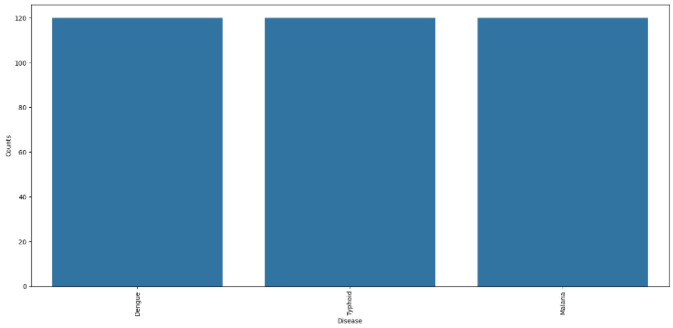

Figure 5. Balance of Diseases

The dataset employed in this project is meticulously organized to facilitate the training of machine learning classifiers. Its equitable distribution guarantees that no particular disease is overrepresented, thereby allowing for the creation of a precise and impartial model for the prediction and diagnosis of dengue, malaria, and typhoid. This balance is crucial for ensuring the model's reliability and its capacity to generalize effectively across the diseases being predicted (dengue, malaria, and typhoid).

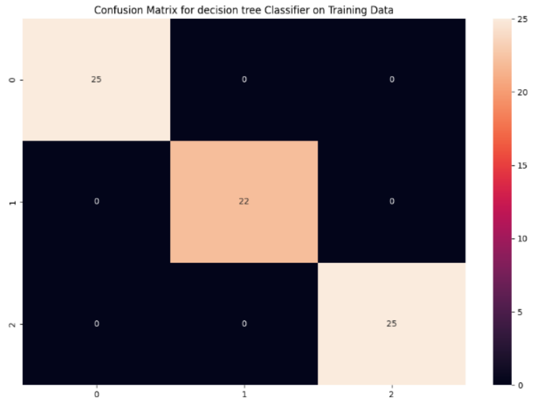

Figure 6. Confusion Matrix for Random Forest Algorithm

This confusion matrix illustrates the efficacy of a Random Forest classifier when applied to test data. Each cell within the matrix denotes the count of predictions corresponding to a particular pairing of actual and predicted classes.

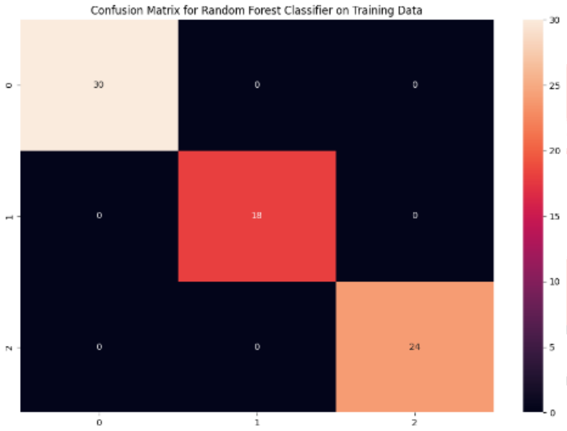

Figure 7. Confusion Matrix for Decision Tree Algorithm

The confusion matrix illustrates the performance of the Decision Tree classifier by juxtaposing the predicted classes against the actual classes across 3 categories (ranging from 0 to 40). The values along the diagonal indicate the instances of correct classifications, also known as True Positives.

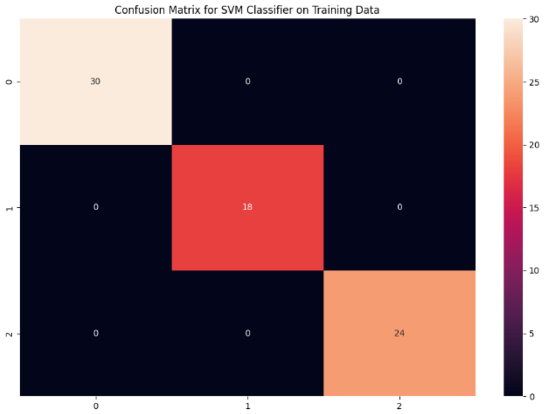

Figure 8. Confusion Matrix for Support Vector Machine

6.1 Accuracy for each Algorithm

Sr.no.	Random forest	Decision Tree	Support vector machine (SVM)
Training	97.03%	90.45%	95.53%
Testing	95.43%	85.77%	93.62%

6.2 Output

```
RF Model Prediction: Typhoid with Probability: 91.00%

SVM Model Prediction: Typhoid with Probability: 54.95%

DT Model Prediction: Typhoid with Probability: 89.52%

Final Prediction: Typhoid

The system predicts you may have Typhoid.
```

7. CONCLUSION

One significant advancement in the application of machine learning technologies in the healthcare industry is the Symptom-based Disease Prediction Project. The project demonstrates how machine learning can reliably predict diseases like dengue, malaria, and typhoid based on patient symptoms by using predictive algorithms including Random Forest, Support Vector Machine, and Decision Tree. These models' exceptional effectiveness, as evidenced by about their high precision rates, highlights capacity to identify the complex relationships between symptoms and related illnesses.

The module created for such a project serves as a foundational illness prediction system. It allows users to enter their symptoms and receive predictions regarding potential diseases, thereby providing a practical tool for early diagnosis. Furthermore, this system holds the potential for future integration into broader platforms, which would enhance its accessibility and usability, ultimately facilitating timely medical responses and improving overall healthcare outcomes.

8. REFERENCES

[1] Sirigineedi, Manikanta & Kumar, Matta & Prakash, Rali & Reddy, Velagala & Tirunagari, Poojitha. (2024). Symptom-Based Disease Prediction: A Machine Learning Approach. Journal of Artificial Intelligence, Machine Learning and Neural Network. 4. 8-17. 10.55529/jaimlnn.43.8.17.

[2] Journal, I. R. J. E. T. (2020). IRJET- Disease Prediction using Machine Learning. IRJET.

[3] Kumar, Anand & U M, Prakash & Sharma, Ganesh. (2021). Disease Prediction and Doctor Recommendation System using Machine Learning Approaches. International Journal for Research in Applied Science and Engineering Technology. 9. 10.22214/ijraset.2021.36234.

[4] Gomathy, C K. (2021). THE PREDICTION OF DISEASE USING MACHINE LEARNING.

[5] Talasila, Bhanuteja & Kolli, Saipoornachand & Kumar, Kilaru & Anudeep, Poonati & Ashish, Chennupati. (2021). Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. International Journal of Innovative Technology and Exploring Engineering. 10. 67-72. 10.35940/ijitee.I9364.0710921.

[6]Jadhav, S., Kasar, R., Lade, N., Patil, M., & Kolte, S. (2019). Disease Prediction by Machine Learning from Healthcare Communities. *International Journal for Research in Applied Science and Engineering Technology*.

[7]Tandon, Sankalp & Chaurasia, Manoj & Singh, Vinit & Kumar, Udit & Kumar, Saurabh. (2023). A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Life. 10.13140/RG.2.2.24229.37601.

[8] Takke, Kunal & Bhaijee, Rameez & Singh, Avanish & Patil, Abhay. (2022). Medical Disease Prediction using Machine Learning Algorithms. International Journal for Research in Applied Science and Engineering Technology. 10. 221-227. 10.22214/ijraset.2022.42135.

[9] Krishnani, Divya & Kumari, Anjali & Dewangan, Akash & Singh, Aditya & Naik, Nenavath. (2019). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms. 367-372. 10.1109/TENCON.2019.8929434.

[10] Yekkala, Indu & Dixit, Sunanda. (2019). Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection: Breakthroughs in Research and Practice. 10.4018/978-1-5225-8185-7.ch011.

[11] Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad, Disease Prediction From Various Symptoms Using Machine Learning (July 27, 2020).

[12] Kosarkar, N., Basuri, P., Karamore, P., Gawali, P., Badole, P., & Jumle, P. (2022). Disease Prediction using Machine Learning. *2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22)*.

[13] Thomas M. Mitchell. 1997. Machine Learning (1st. ed.). McGraw-Hill, Inc., USA.

[14] Jakkula, Vikramaditya R.. “Tutorial on Support Vector Machine (SVM).” (2011).

[15] Sirigineedi, Manikanta & Kumar, Matta & Prakash, Rali & Reddy, Velagala & Tirunagari, Poojitha. (2024). Symptom-Based Disease Prediction: A Machine Learning Approach. Journal of Artificial Intelligence, Machine Learning and Neural Network. 4. 8-17. 10.55529/jaimlnn.43.8.17.