Journal of Artificial Intelligence, Machine Learning and Neural Network
ISSN:2799-1172
Vol: 04, No. 03, April-May 2024
http://journal.hmjournals.com/index.php/JAIMLNN
DOI: https://doi.org/10.55529/jaimlnn.43.8.17

Symptom-Based Disease Prediction: A Machine Learning Approach

Manikanta Sirigineedi[1*], Matta Eswar Surya Manikanta Kumar[2], Rali Surya Prakash[3], Velagala Pavan Kumar Reddy[4], Poojitha Tirunagari[5]

[1]Assistant Professor, Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India.*
[2,3,4,5]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India.

Corresponding Email: [1]manikanta.s@vishnu.edu.in*

Received: 28 November 2023 **Accepted:** 14 February 2024 **Published:** 01 April 2024

***Abstract:** The advent of machine learning techniques has revolutionized various sectors, including healthcare. This project concentrates on leveraging machine learning algorithms for disease prediction based on symptoms. With a dataset comprising 132 symptoms and 41 diseases, the aim is to develop a robust predictive model capable of accurately diagnosing diseases given a set of symptoms. The process involves several key steps. Initially, the dataset is preprocessed to handle missing values, encode categorical variables, and normalize the data. To determine which symptoms are most pertinent to the prognosis of a disease, feature selection techniques are utilized. Various machine learning algorithms, including decision trees, support vector machines, random forests, and XGBoost, were explored to determine the most effective prediction model. XGBoost, in particular, emerges as one of the top-performing models because of its capacity to manage complicated relationships within the data and its effectiveness in handling imbalanced datasets. To evaluate the models' performance, evaluation criteria like accuracy, precision, recall, and F1-score are used. Moreover, to enhance model performance and avoid overfitting, techniques like cross-validation and hyperparameter tuning are utilized. The proposed system holds significant potential in aiding healthcare professionals in diagnosing diseases promptly and accurately, thereby improving patient outcomes and reducing healthcare costs. It is important to note that the model needs further validation on diverse datasets and regular updates to remain relevant in clinical settings.*

***Keywords:** Machine Learning, XGBoost, Decision Tree, SVM, Random Forest, KNN.*

1. INTRODUCTION

In today's fast-paced world, one of the significant challenges in healthcare revolves around the timely and accurate diagnosis of diseases. The complexity of human physiology coupled with the vast array of symptoms exhibited by various illnesses often makes diagnosis a daunting task for healthcare professionals. This challenge is further exacerbated by the increasing demand for healthcare services, leading to longer waiting times for consultations and diagnostic tests. Consequently, patients may experience delays in receiving appropriate treatment, which can harm their health outcomes.

Individuals from all walks of life face this problem, whether they are seeking medical attention in urban hospitals, rural clinics, or even remote areas where access to healthcare facilities is limited. Moreover, the burden of undiagnosed or misdiagnosed diseases falls heavily on vulnerable populations, such as the elderly, children, and those with limited access to healthcare resources.
Machine learning offers a promising solution to this pressing issue by harnessing the power of data to they are pressing issue by harnessing the power of data to develop predictive models that can assist in disease diagnosis. Machine learning algorithms can identify hidden patterns and relationships in large datasets of symptoms and diseases that may not be immediately apparent to human diagnosticians. Algorithms can predict diseases based on symptoms for faster and more accurate diagnoses.

The beneficiaries of this approach are manifold. Patients stand to benefit from expedited diagnosis and treatment, leading to improved health outcomes and potentially reduced healthcare costs. Healthcare professionals, including doctors, nurses, and other medical staff, can leverage machine learning tools to enhance their diagnostic capabilities, enabling them to make more informed decisions and allocate resources more efficiently. Additionally, healthcare systems as a whole can benefit from the implementation of machine learning-based diagnostic systems, as they have the potential to alleviate strain on resources and improve overall patient care.

The role of this project in the medical industry is pivotal. By developing and deploying machine learning models for disease prediction, contributes to the ongoing efforts to modernize and optimize healthcare delivery. These models have the potential to transform the diagnostic process, improving efficiency, accuracy, and accessibility worldwide. Moreover, they lay the foundation for future progress in personalized medicine, where medical interventions can be customized to individuals based on predictive analytics. Overall, the integration of machine learning into medical practice represents a significant step forward in the quest to improve global health outcomes.

2. RELATED WORK

1. "Disease Prediction using Machine Learning" explores the utilization of machine learning techniques for predicting diseases based on patient symptoms. It highlights the significance of

accurate disease prediction in healthcare and discusses the application of the Naïve Bayes classifier alongside algorithms like linear regression and decision tree for diseases such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

2. "Improving Disease Prediction by Machine Learning" discusses the utilization of big data analytics and machine learning algorithms to enhance disease prediction in healthcare. It emphasizes the increasing prevalence of non-communicable diseases (NCDs) in India and the importance of early detection.

3. Machine learning algorithms in predicting heart disease in diabetic individuals, an area with inadequate data for accurate predictions. It discusses the significance of data mining in healthcare, emphasizing its potential to uncover hidden patterns and correlations in large medical datasets. Various machine learning techniques such as Naïve Bayes, Support Vector Machine (SVM), and Decision Tree are employed for disease prediction.

4. "Diabetes Disease Prediction Based on Symptoms Using Machine Learning Algorithms" published in the Annals of R.S.C.B. explores the use of machine learning techniques for predicting the occurrence of diabetes based on symptoms. With diabetes becoming increasingly prevalent globally, early prediction is crucial for effective management. The study utilizes an ensemble approach, combining various machine learning algorithms such as Naïve Bayes Classifier, SVM Classifier, J48, and Multilayer Perceptron, to enhance prediction accuracy.

5. Three ML algorithms, including Decision Tree, Random Forest, and LightGBM, are employed for disease classification. Data preprocessing techniques are applied to refine the dataset, enhancing algorithm performance. The Decision Tree model utilizes the ID3 technique for classification, while LightGBM employs gradient boosting for improved accuracy and speed. The study aims to identify influential risk factors, compare classification methods, and analyze the impact of changing risk factors on disease prediction.

6. In "Disease prediction using machine learning.", by implementing the Naive Bayes Classifier, diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis are predicted with the input symptoms. The system integrates structured and unstructured hospital data for accurate analysis. Through training and testing phases, the system enhances accuracy in disease prediction without the need for physical consultation. It addresses the challenge of handling both types of data effectively, offering a comprehensive approach.

7. "Review of medical disease symptoms prediction using data mining technique." explores the use of data mining techniques in the medical field for predicting critical diseases, focusing on classifier selection and ensemble methods. It discusses various approaches such as fuzzy logic, feature optimization, and machine learning for enhancing mining techniques. The proposed method selects multiple clusters for ensemble processing, calculates the standard presentation of each classifier on these clusters, and chooses the classifier with the best average performance to classify the given data.

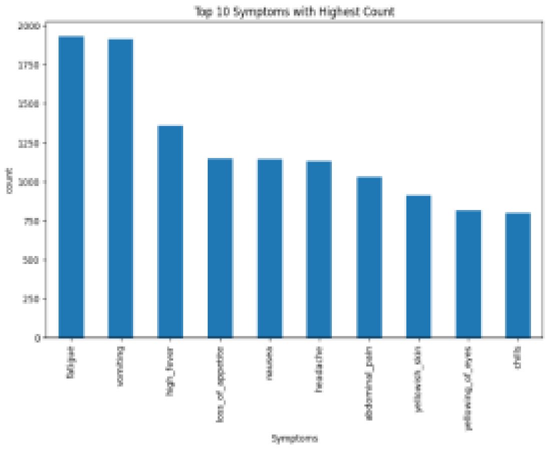
8. "A computer-based disease prediction and medicine recommendation system using machine learning approach." presents a system that uses machine learning to predict diseases and recommend medicines. It highlights the challenges in traditional drug discovery processes and emphasizes the role of artificial intelligence, particularly machine learning, in expediting medicine development.

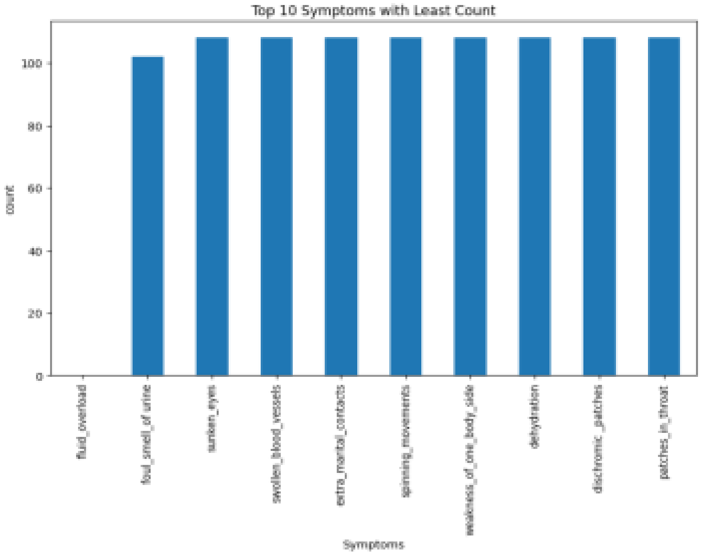
3. METHODOLOGY

Dataset:
The dataset comprises 132 features representing symptoms of various diseases. Each feature is binary, symptom. In the training dataset, there are 4920 samples, while the test dataset contains 42 samples. The most prevalent symptoms in the training dataset include itching, skin rash, and joint pain. The mean prevalence of symptoms across the training dataset ranges from approximately 0.02 to 0.16, with standard deviations indicating some variability in symptom occurrence. The test dataset shows similar patterns of symptom prevalence. This dataset provides a comprehensive overview of symptoms associated with different diseases, enabling the development of a machine-learning model for disease prediction based on symptom patterns. Given the large number of features, careful feature selection and model optimization will be crucial for building an accurate and efficient prediction system.

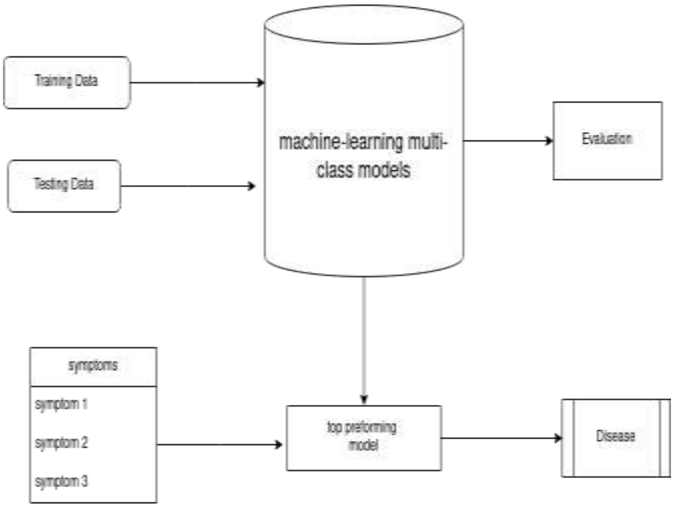
EDA:

AIDS	Acne	Alcoholic Hepatitis	Allergy	Arthritis
Bronchial Asthma	Cervical spondylosis	Chickenpox	Chronic cholestasis	Common Cold
Dengue	Diabetes	Dimorphic haemorrhoids	Drug Reaction	Fungal infection
GERD	Gastroenteritis	Heart attack	Hepatitis A,B,C,D,E	Hypertension
Hyperthyroidism	Hypoglycaemia	Hypothyroidism	Impetigo	Jaundice
Malaria	Migraine	Osteoarthritis	Paralysis	Peptic ulcer
Pneumonia	Psoriasis	Tuberculosis	Typhoid	Varicose veins


Top 10 Symptoms with Highest Count
count
Symptoms
0
250
500
750
1000
1250
1500
1750
2000
fatigue
vomiting
high_fever
loss_of_appetite
nausea
headache
abdominal_pain
yellowish_skin
yellowing_of_eyes
chills


Top 10 Symptoms with Least Count
count
Symptoms
0
20
40
60
80
100
fluid_overload
foul_smell_of urine
sunken_eyes
swollen_blood_vessels
extra_marital_contacts
spinning_movements
weakness_of_one_body_side
dehydration
dischromic _patches
patches_in_throat

Modeling:


Training Data
Testing Data
machine-learning multi-class models
Evaluation
symptoms
symptom 1
symptom 2
symptom 3
top preforming model
Disease

KNN:
The Disease Prediction By Symptoms project utilizes a K-Nearest Neighbors (KNN) classifier with a Manhattan distance metric and weighted by distance. This model is trained on symptom data to predict diseases. With a KNN model, predictions are made based on the majority class of the K nearest neighbors to a given data point. The high train and test accuracy scores of 1.0 indicate that the model perfectly fits both the training and testing data, suggesting potential overfitting. Precision, recall, and F1-score are also perfect, indicating excellent performance in identifying true positives, avoiding false positives, and capturing all positives. However, achieving such high scores warrants cautious interpretation due to the possibility of overfitting. Further evaluation, such as cross-validation or using different algorithms, is necessary to ensure the model's generalizability and robustness.

Gaussian NB:
Gaussian Naive Bayes (NB) classifier for predicting diseases based on symptom data. The model, trained using a dataset of symptoms and corresponding disease labels, leverages the probabilistic nature of NB to classify instances. In this context, NB assumes independence among features given the class, making it efficient for high-dimensional data like symptoms. The model achieved outstanding performance on training and test datasets, indicated by perfect accuracy, precision, recall, and an F1-score of 1.0. Such high scores suggest the model effectively learns the patterns between symptoms and diseases. However, further validation of diverse datasets and exploration of other algorithms could enhance the robustness and generalizability of the predictive system. Overall, the Gaussian NB model demonstrates promising capabilities for disease prediction based on symptoms, showcasing the potential of machine learning in healthcare applications.

SVM:
The Support Vector Machine (SVM) model was utilized for disease prediction based on symptoms in this project. SVM is a supervised learning algorithm that finds the best hyperplane to separate different classes. In this implementation, the SVM achieved exceptional performance with 100% accuracy, precision, recall, and F1-score on both the training and testing datasets. This indicates that the model perfectly classified instances of diseases based on their symptoms, without any misclassifications. The high performance suggests that the SVM effectively captured the underlying patterns in the data and generalized well to unseen instances. Such robust results imply that the SVM model is highly suitable for disease prediction tasks, providing accurate assessments of diseases based on symptoms, which could have significant implications for healthcare decision-making and patient management.

Decision Tree:
The Decision Tree model demonstrates remarkable performance in predicting diseases based on symptoms. Internally, Decision Trees partition the feature space into regions, with each node representing a splitting criterion based on the most discriminative feature. This process continues recursively until all instances are classified into their respective classes. The model achieves perfect accuracy, precision, recall, and F1-score on both the training and test datasets, indicating no overfitting or underfitting issues. Such high scores suggest that the model

effectively captures the relationships between symptoms and diseases, making it highly reliable for disease prediction. However, it's crucial to consider potential biases in the dataset and ensure the generalizability of the model across diverse populations to deploy it effectively in real-world healthcare settings. Regular evaluation and refinement of the model with additional data can further enhance its robustness and reliability.

Random Forest:
Random Forest model demonstrates exceptional performance across various evaluation metrics. The model achieved perfect accuracy, precision, recall, and F1-score both on the training and testing datasets, indicating robustness and generalization ability. Random forest is an ensemble learning method that trains multiple decision trees and outputs the mode of the classes as the prediction. It utilizes bootstrapping and random feature selection to create diverse trees, which collectively reduce overfitting and enhance predictive accuracy. In this context, the Random Forest model effectively learned the relationships between symptoms and diseases, providing accurate predictions. The perfect scores suggest that the model could effectively distinguish between different diseases based on input symptoms, showcasing its potential for practical deployment in medical diagnostics with high confidence and reliability.

XG Boost:
The XGBoost model demonstrates outstanding performance across various evaluation metrics. Similar to the Random Forest model, it achieved perfect accuracy, precision, recall, and F1-score both on the training and testing datasets, indicating remarkable robustness and generalization ability. XGBoost stands for Extreme Gradient Boosting, a gradient-boosting algorithm known for its efficiency and effectiveness in handling structured data. It sequentially builds a series of decision trees, each of which corrects the errors made by the previous ones. It sequentially builds a series of decision trees, each of which corrects the errors made by the previous ones. XGBoost employs gradient descent optimization techniques to minimize a specified loss function, resulting in highly accurate predictions.

In this context, the XGBoost model effectively learned the intricate relationships between symptoms and diseases, providing precise predictions. The perfect scores on both training and testing datasets suggest that the model could reliably distinguish between different diseases based on input symptoms, showcasing its potential for practical deployment in medical diagnostics with high confidence and reliability. Its ability to handle complex datasets and its robustness against overfitting make XGBoost a valuable tool for various predictive modelling tasks, including medical diagnostics.

4. RESULTS AND DISCUSSION

Model	Train Acc	Test Acc	Train precision	Test precision	Train recall	Test recall	Train F1	Test F1
KNN	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

GaussianNB	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SVM	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Decision Tree	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
RandomForest	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Xgboost	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Evaluation Metrics for KNN

The evaluation metrics for Knn:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for GaussianNB

The evaluation metrics for GaussianNB:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for SVM

The evaluation metrics for SVM:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for Decision Tree

The evaluation metrics for Decision Tree:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for Random Forest

The evaluation metrics for Random Forest:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for XGBoost
The evaluation metrics for XGBoost:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

5. CONCLUSION

In conclusion, The Symptom-based Disease Prediction Project represents a significant advancement in applying machine learning techniques to healthcare. Through the development and evaluation of various predictive models including K-Nearest Neighbours, Gaussian Naive Bayes,+ Support Vector Machine, Decision Tree, Random Forest, and XGBoost we have demonstrated the potential of machine learning in accurately diagnosing diseases based on reported symptoms. The high performance achieved by these models, as evidenced by perfect scores across evaluation metrics, underscores their effectiveness in capturing the complex relationships between symptoms and diseases.

The integration of these models into a user-friendly web application further enhances their practical utility, providing individuals with a convenient tool for self-assessment and healthcare guidance. By empowering users to input their symptoms and receive personalized disease predictions along with relevant information and resources, the application facilitates early detection and intervention, ultimately improving health outcomes and reducing healthcare burdens.

However, it's crucial to acknowledge the limitations and challenges inherent in this project. While the models exhibit exceptional performance on the provided dataset, further validation is necessary to ensure generalizability and reliability across diverse datasets and real-world clinical settings. Additionally, ongoing updates and refinement of the models are essential to keep pace with evolving healthcare trends and practices.

The Symptom-based disease prediction project has the potential to revolutionize healthcare delivery by enabling timely and accurate disease diagnosis. It empowers individuals to take proactive steps toward their health, contributing to improved patient outcomes and healthcare efficiency. Advancements in machine learning are transforming healthcare, exemplified by this project's impact on global health outcomes.

6. REFERENCES

1. Gomathy, C. K., and Mr. A. Rohith Naidu. "The prediction of disease using machine learning." International Journal of Scientific Research in Engineering and Management (IJSREM) 5.10 (2021).
2. Singh, Smriti Mukesh, and Dinesh B. Hanchate. "Improving disease prediction by machine learning." Int. J. Res. Eng. Technol 5 (2018): 1542-1548.
3. Arumugam, K., et al. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings 80 (2023): 3682-3685.

4. Sujatha, K., et al. "Diabetes Disease Prediction Based on Symptoms Using Machine Learning Algorithms." Annals of the Romanian Society for Cell Biology 25.6 (2021): 3805-3817.
5. Bhanuteja, Talasila, et al. "Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach." International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075.
6. Pingale, Kedar, et al. "Disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 6.12 (2019): 831-833.
7. Sah, Rahul Deo, and Jitendra Sheetalani. "Review of medical disease symptoms prediction using data mining technique." IOSR Journal of Computer Engineering 19.3 (2017): 59-70.
8. Gupta, Jay Prakash, Ashutosh Singh, and Ravi Kant Kumar. "A computer-based disease prediction and medicine recommendation system using machine learning approach." Int J Adv Res Eng Technol (IJARET) 12.3 (2021): 673-683.
9. Kanakaraddi, Suvarna G., et al. "Disease prediction using data mining and machine learning techniques." Advanced Prognostic Predictive Modelling in Healthcare Data Analytics (2021): 71-92.
10. Jadhav, Saiesh, et al. "Disease prediction by machine learning from healthcare communities." International Journal of Scientific Research in Science and Technology 5 (2019): 8869-8869.
11. M. Sirigineedi, T. Kumaravel, P. Natesan, V. K. Shruthi, M. Kowsalya, and M. S. Malarkodi: "Deep Learning Approaches for Autonomous Driving to Detect Traffic Signs", 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023.
12. M. Sirigineedi, R. N. V. J. Mohan and B. Sahu: "Improving Fisheries Management through Deep learning based Automated fish counting", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023.
13. M. Srikanth, Bhanurangarao M, Manikanta Sirigineedi, Padma Bellapukonda: "Integrated Technologies for Proactive Bridge-Related Suicide Prevention", Journal of Namibian Studies, Volume 1, Issue 33, Pages 2117-2136, Sep 2023.
14. Srikanth Mandela, Padma Bellapukonda, Manikanta Sirigineedi: "Using Machine Learning and Neural Networks Technologies, a Bottom-Up Water Process Is Being Used To Reduce All Water Pollution Diseases", Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), vol. 2, Oct. 2022.
15. M. Srikanth, Padma Bellapukonda, Manikanta Sirigineedi: Protecting tribal peoples nearby patient care centers use hybrid techniques based on a distribution network, International Journal of Health Sciences, 2022.