iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49641

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

Disease Prediction and Treatment Recommendation Using Machine Learning

Saurav Mahata[1], Yash Bhaveshbhai Kapadiya[2], Vishal Kushwaha[3], Vatsal Joshi[4], Yassir Farooqui[5]
B.Tech CSE, Parul Institute of Engineering and Technology, Vadodara

Abstract: ***Despite the availability of advanced technology and easy access to information, many people still rely on traditional methods of seeking medical treatment, such as visiting hospitals and consulting doctors for even minor symptoms. However, this approach can be time-consuming and inefficient, as patients with minor illnesses can take up valuable resources that could be better used to treat more serious cases. As a result, this research proposes a new approach to disease prediction using machine learning and symptom-based analysis. The goal is to develop a predictive model that can accurately identify potential diseases based on a patient's symptoms. This model utilizes machine learning techniques to analyze and process symptom data, allowing for quick and precise disease prediction. The study uses a large dataset of patient symptoms and medical records to train and test the model, which demonstrated high accuracy in predicting diseases. The results of this study suggest that the proposed model could be a useful tool for early diagnosis and treatment of diseases, with the potential to improve healthcare outcomes. Overall, this research represents an important contribution to the field of healthcare informatics, with possible applications in disease prevention, treatment, and management.***

Keywords: Machine Learning, Dataset, KNN, Python-Flask

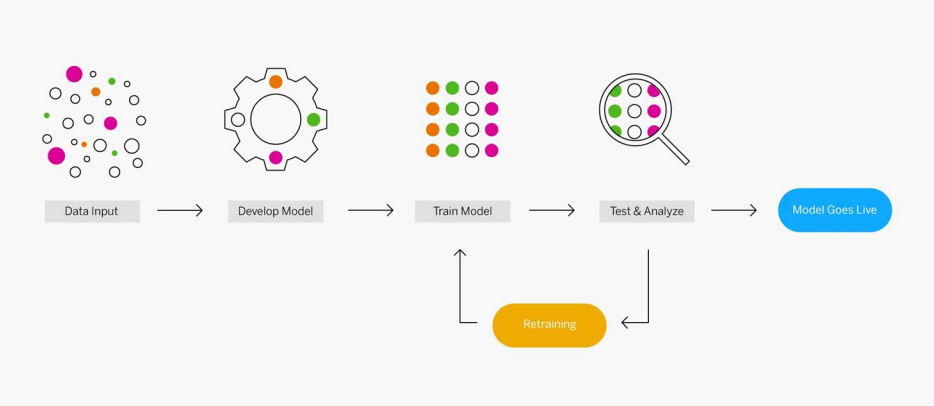
I. INTRODUCTION

Machine learning has gained significant attention in recent years due to its potential to analyze large volumes of data and provide insights that were previously unattainable. The healthcare industry has also recognized the value of machine learning in the diagnosis, treatment, and management of various diseases. Machine learning algorithms have the ability to learn from data and provide accurate predictions, which can significantly improve healthcare outcomes. In particular, the use of machine learning in disease prediction has gained widespread attention in the healthcare industry. This paper aims to discuss the application of machine learning algorithms in disease prediction, with a focus on the k-nearest neighbors (KNN) algorithm.

The KNN algorithm is a type of supervised learning algorithm that is widely used in machine learning for classification and regression analysis. The algorithm is based on the principle of similarity, where it searches for the k nearest neighbors to a given data point and predicts the class or value based on the most common class or the average value of the k nearest neighbors. In the case of disease prediction, the KNN algorithm can be used to predict the disease based on the similarity of symptoms between the patient and a dataset of known cases. The algorithm takes into account the patient's symptoms and compares them with the symptoms of previously diagnosed patients to determine the most likely disease.

The KNN algorithm has several advantages in disease prediction, including its simplicity, speed, and accuracy. It does not require any assumptions about the underlying distribution of data, and it can handle noisy and incomplete data. Additionally, the KNN algorithm is highly interpretable, which makes it easier for physicians to understand the predictions and make informed decisions.

Disease prediction is a critical aspect of modern healthcare, enabling early diagnosis and prompt treatment that can improve patient outcomes and save lives. Traditional approaches to disease prediction rely on the expertise of medical professionals and diagnostic tests, which can be time-consuming, expensive, and often limited by the availability of trained specialists. However, recent advances in machine learning have opened up new possibilities for accurate and efficient disease prediction using large datasets of patient data. By applying machine learning algorithms to patient data, it is possible to identify complex patterns and relationships that may not be immediately apparent to human experts. This can lead to more precise and timely diagnosis of diseases, as well as an improved understanding of disease risk factors and treatment outcomes. In this context, this paper presents a review of the current state-of-the-art in machine learning-based disease prediction, including a discussion of the challenges, opportunities, and future directions for research in this field. The paper also presents several case studies that demonstrate the potential of machine learning in disease prediction, with a focus on applications in clinical decision-making, personalized medicine, and public health. Overall, this paper aims to provide a comprehensive overview of the use of machine learning in disease prediction and to highlight its potential for transforming the future of healthcare.



II. LITERATURE SURVEY

Kohli, P.,Arora, H. [1] discusses the use of various classification algorithms to diagnose three different diseases (Heart Disease, Breast Cancer, and Diabetes) based on data from the UCI repository. The proposed method had high predictive accuracy, with 87.1% for Heart Disease Detection using Logistic Regression, 85.71% for Diabetes Predictability using a Vector Support Machine (line kernel), and 98.57% for AdaBoost Cancer Screening.

Grampurohit, S. & Sagarnal, C. [2] analyzes performance metrics for various types of machine learning models used in diagnostic tests. Naive Bayes, Decision Trees, and K-Nearest Neighbor are commonly used, but the Support Vector Machine (SVM) is the most effective in diagnosing kidney disease and Parkinson's disease, while Logistic Regression (LR) plays a key role in predicting heart disease. Random Forest and Convolutional Neural Networks (CNN) are accurate in predicting breast and common diseases, respectively.

Kanchan, B. & Kishore, M. [3] aims to improve the accuracy of machine learning algorithms by identifying the minimum number of attributes needed to predict heart disease. The study compares the accuracy of various algorithms using classifier accuracy, time to build a model, mean total error, and ROC location. The results show that using PCA to reduce the number of attributes in the database improves the performance of SVM, Naive Bayes, and Decision Tree algorithms in predicting both heart disease and diabetes.

Mathew, R., Varghese, S., Joy, S. & Alex, S. [4] proposes developing a chatbot app that uses natural language processing and machine learning to interact with users and identify their symptoms, predict diseases, and recommend treatments. This chatbot app can be used for daily checkups and can help people become more aware of their health status, encouraging them to take the necessary steps to stay healthy.

Dahiwade, D., Patle, G. & Meshram, E.[19] uses K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) learning algorithms to predict common diseases based on patient symptoms, life habits, and diagnostic information. The accuracy of the CNN algorithm in predicting common diseases was found to be 84.5%, higher than that of the KNN algorithm.

Ambekar, S. , Phalnikar,R. , [20] discusses the importance of accurate data analysis to diagnose diseases and provide early patient care. The study aimed to predict heart disease using the Naïve Bayes and KNN algorithms and proposed expanding this work to predict disease risk using systematic data.

III. RESEARCH FINDINGS

The use of machine learning in disease prediction and diagnosis. discusses the use of various classification algorithms to diagnose Heart Disease, Breast Cancer, and Diabetes based on data from the UCI repository. analyzes the performance metrics for various types of machine learning models used in diagnostic tests and concludes that Support Vector Machine (SVM) is the most effective in diagnosing kidney disease and Parkinson's disease, while Logistic Regression (LR) plays a key role in predicting heart disease. Paper 3 aims to improve the accuracy of machine learning algorithms by identifying the minimum number of attributes needed to predict heart disease and compares the accuracy of various algorithms using classifier accuracy, time to build a model, mean total error, and ROC location. machine learning to interact with users and identify their symptoms, predict diseases, and recommend treatments. Finally, Paper [19] uses K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) learning algorithms to predict common diseases based on patient symptoms, life habits, and diagnostic information, while Paper[20] emphasizes the importance of accurate data analysis to diagnose diseases and provide early patient care.

IV. PROPOSED METHODOLOGY

Initially we take disease dataset from Kaggle website and that is in the form of disease list with its symptoms. After that preprocessing is performed on that dataset for cleaning that is removing comma, punctuations and white places. And that is used as training dataset. After that feature extracted and selected. Then we classify that data using multiple classification techniques . Based on machine learning we can predict accurate disease.

In this project seven machine-learning algorithms namely Logistic Regression, Random Forest classification, XGB (Extreme Gradient Boosting) classification, KNN (K-nearest Neighbors) classification, Decision Tree classification, Naïve Bayes, and SVC (Support Vector Machine)techniques are applied on diseases data set For the implementation of any algorithm or techniques, in general, there are some certain steps need to be considered.

1) Data Collection Phase: It is the first and very important phase from where the exact process begins. We need to collect the related data, which suits the requirement.

2) Data Validation Phase: In this phase, we will check whether the data we have collected will exactly be related to our application.

3) Data Analysis Phase: In this phase, we analyze the collected data by implementing it in various models with various criteria's. 4) Data Reporting: It involves the reporting of data, which have been collected from the previous stage. For Data collection we are using a publicly available dataset downloaded from Kaggle named sympthoms.csv.

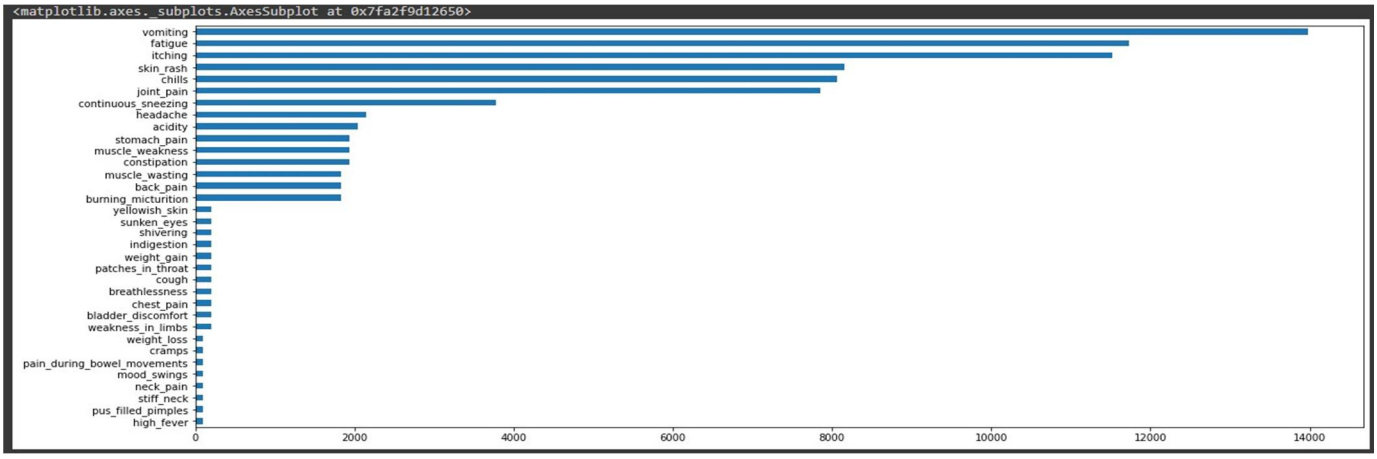
A. About Dataset

The dataset includes the symptoms of multiple diseases. The dataset is highly imbalanced. The dataset contains multiple value fields. There where columns like Disease,Symptom1, Symptom2, Symptom3 upto Symptom17 and the dataset also cantain multiple nun values initially.there where total 4920 rows and 18 columns in dataset.

```
[ ]  df.head()
```

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Sy
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN	NaN	NaN	NaN	
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN	NaN	NaN	NaN	

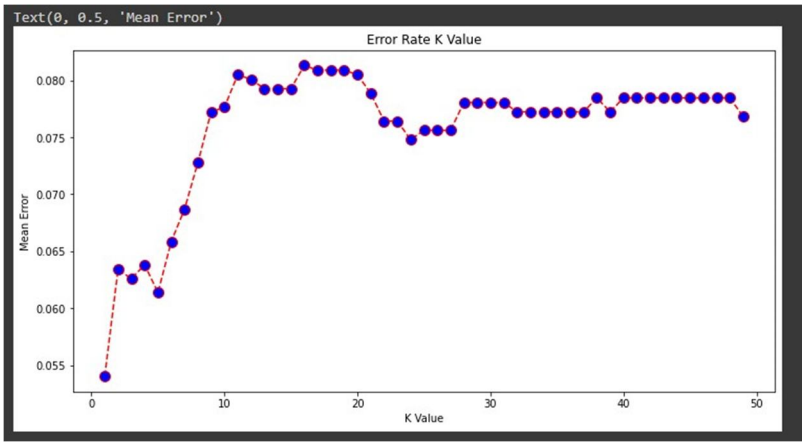
Therefore, had to reframe the dataset into Boolean form by replacing the symptoms name columns name on the bases of the occurrence of symptoms on the dataset. after being converted into a numerical value. Then classify symptoms and diseases as per there uniqueness of occurrence.



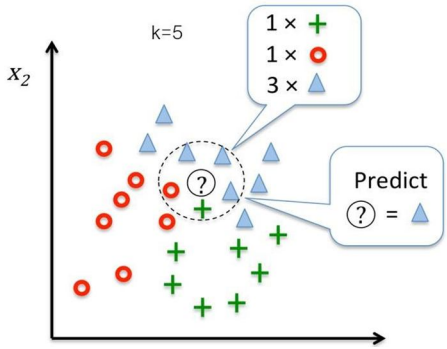
B. Implementation Of Algorithms

Using multiple Algorithms to find the best suited as per our dataset. following are the algorithms we have tried.

1) *Logistic Regression:* Logistic regression is a method for predicting the result of a dependent variable that has categories. This means that the result must be a discrete or categorical value, such as Yes or No, 0 or 1, or true or false. However, instead of providing an exact value of either 0 or 1, logistic regression produces probability values that range from 0 to 1.
2) *Random Forest Algorithm:* A random forest consists of multiple decision trees that are built on different subsets of the dataset. By taking the average of the predictions from these trees, the random forest aims to improve the accuracy of its predictions. Unlike a single decision tree, the random forest uses a voting system to determine the final output, where the prediction with the majority of votes from the individual trees is selected.
3) *XGBoost:* This AI technique is used in tasks such as classification and regression, among others. It creates a model of predictions by combining a set of weaker prediction models, which are often referred to as decision trees.
4) *Decision Tree Classification:* In a decision tree, there are two types of nodes: Decision Nodes and Leaf Nodes. Decision Nodes are utilized to make decisions and contain multiple branches, while Leaf Nodes represent the output of those decisions and do not contain any further branches. A visual way of finding all the feasible solutions to a problem or decision, taking into account specific conditions, is known as a graphical representation.
5) *Naïve Bayes Classifier:* The Naïve Bayes Classifier is an uncomplicated but highly efficient classification algorithm that can be used to create rapid machine learning models capable of swiftly making predictions.
6) *Support Vector Machine:* The objective of the SVM algorithm is to establish an optimal line or decision boundary, known as a hyperplane, that can effectively separate n-dimensional space into different classes. This enables easy classification of new data points in the future, placing them accurately into the correct category.
7) *K-NN:* The K-NN algorithm stores all of the available data and classifies a new data point by measuring its similarity to the existing data. This allows new data to be easily categorized into an appropriate group using the K-NN algorithm. Although the K-NN algorithm can be utilized for both classification and regression, it is primarily applied to classification problems.



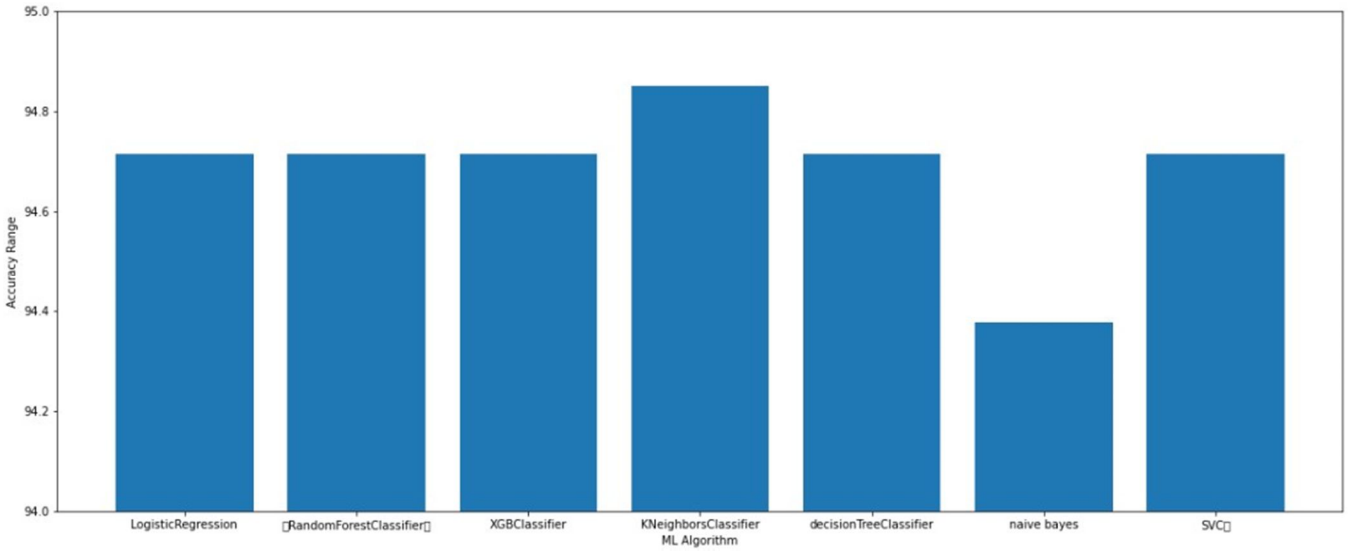
As from the research work we come to know from all seven algorithms KNN is the best suited for this dataset as we are getting the best K value for the dataset.



V. RESULT ANALYSIS

The study used seven different machine-learning algorithms to detect symptoms in various diseases. The results showed that the KNN (K-nearest Neighbors) algorithm had the highest accuracy and the lowest error rate. A graphical representation of these findings is shown in Figure. The table presents a comparison of accuracy statistics for Logistic Regression, Random Forest classification, XGB (Extreme Gradient Boosting) classification, KNN (K-nearest Neighbors) classification, Decision Tree classification, Naïve Bayes, and SVC (Support Vector Machine) that were tested for different partition.

Sr. No	Algorithms	Accuracy score	Cross value score
1	Logistic Regression,	94.715447	95.101626
2	Random Forest classification	94.715447	95.101626
3	XGB (Extreme Gradient Boosting) classification	94.715447	95.101626
4	KNN (K-nearest Neighbors) classification	94.850949	94.695122
5	Decision Tree classification	94.715447	95.101626
6	Naïve Bayes	94.376694	95.101626
7	SVC (Support Vector Machine)	94.715447	93.821138



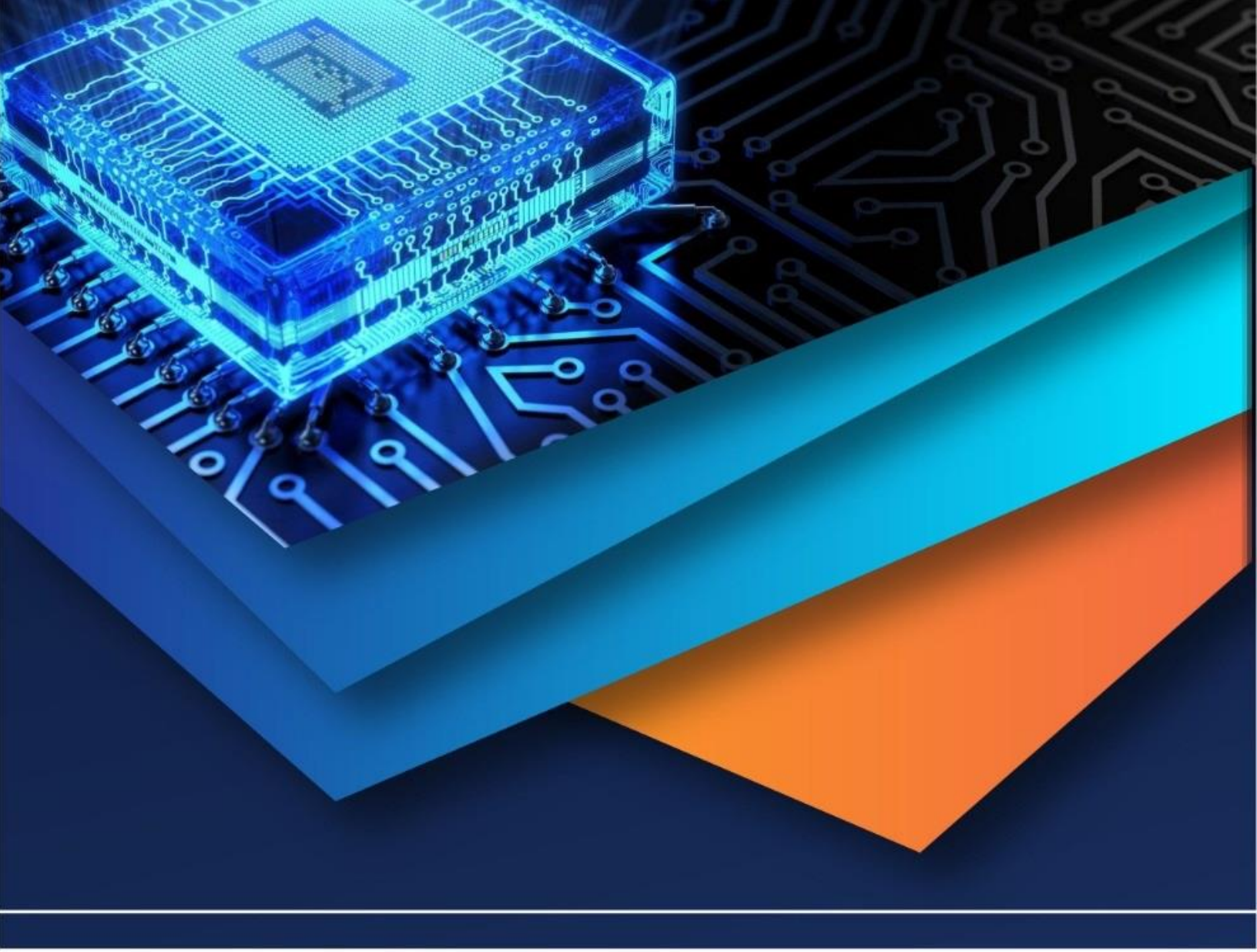
VI. CONCLUSION

This paper describes the use of machine learning algorithms to predict the occurrence of diseases based on their symptoms. The model was trained using different training and testing data units, and it was found that the KNN (K-nearest Neighbors) algorithm performed the best in detecting diseases based on symptoms. The experiment was conducted using varying training and testing units, and the results showed that KNN was the most effective algorithm for this task.

REFERENCES

[1] Kohli, P.,Arora, H.,2018, Application of Machine Learning in Disease Prediction, .In: 2018 4th International Conference on Computing Communication and Automation (ICCCA) , Dec. 14-15 , 2018. Greater Noida, India
[2] Grampurohit, S. & Sagarnal, C., 2020. Disease Prediction using Machine Learning Algorithms. In: 2020 International Conference for Emerging Technology (INCET), Jun 5-7, 2020, Belgaum, India
[3] Kanchan, B. & Kishore, M., 2016. Study of machine learning algorithms for special disease prediction using principal of component analysis. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC),Dec 22-24, 2016, Jalgaon, India.
[4] Mathew, R., Varghese, S., Joy, S. & Alex, S.,2019. Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning. In: Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), April 23-25, 2019, Tirunelveli, India.
[5] Grampurohit, S.,Sagarnal,C., 2020 Disease Prediction using Machine Learning Algorithms, in 2020 International Conference for Emerging Technology (INCET), 5-7 June 2020, Belgaum, India
[6] Skrebeca ,J. ,Kalniete, P., Goldbergs, J. , Pitkevica, L., Tihomirova, D. , & Romanovs, A.2021. Modern Development Trends of Chatbots Using Artificial Intelligence (AI). In: 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University. 14-15 Oct. 2021. Riga,Latvia.

[7] Karayılan,T., Kılıç, O., 2017. Prediction of heart disease using neural network.In 2017 International Conference on Computer Science and Engineering (UBMK). 5-8 Oct. 2017, Antalya, Turkey
[8] Kandpal, P. , Jasnani, K. , Raut, R. & Bhorge,S. 2020, Contextual Chatbot for Healthcare Purposes (using Deep Learning). In: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). 27-28 July 2018. London,UK.
[9] Kulothunkan, P. , Mohd, N. , Khairul , A., 2021, SEQ2SEQ++: A Multitasking-Based Seq2seq Model to Generate Meaningful and Relevant Answers,in IEEE Access (Volume: 9), 06 December 2021 , IEEE, pp 164949 – 164975
[10] Jha,P. , Biswas, T., Utkarsha, S., Ahuja,K.,2021, Prediction with ML paradigm in Healthcare System,in 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 4-6 Aug. 2021, Coimbatore, India
[11] Ju, L., 2022. Improving Medical Images Classification with Label Noise Using Dual-Uncertainty Estimation. In: IEEE Transactions on Medical Imaging, Jan 07, 2022.
[12] Mekha, Panuwat. & Teeyasuksaet, Nutnicha.,2021. Image Classification of Rice Leaf Diseases Using Random Forest Algorithm. In: 6th International Conference on Digital Arts, Media and Technology (DAMT) and 4th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (NCON),3-6 March 2021, Cha-am, Thailand.
[13] Mohammad,I.,Swathi,T. Kireeti,M., Reddy,Y. Kishore. & Lakshmi,K.Naga.,2019. Design and Implementation of Student Chat Bot using AIML and LSA. In: International Journal of Innovative Technology and Exploring Engineering (IJITEE) , 6 April 2019,
[14] Mishra, S., Zhang, Y., Chen, D. & Hu, X., 2021. Data-Driven Deep Supervision for Medical Image Segmentation. In: IEEE Transactions on Medical Imaging, Jan 14, 2021.
[15] Daniel, Gwendal. & Cabot, Jordi., 2021. The Software Challenges of Building Smart Chatbots.In: 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 25-28 May 2021,Madrid,ES.
[16] Gupta, J. Singh, V. & Kmar, I., 2021. Florence- A Health Care Chatbot.In:2021 7th International Conference On Advanced Computing And Communication Systems (ICACCS) , March 19-20, 2021. Coimbatore. India.
[17] Purushottam; Kanak Saxena; Richa Sharma,2015, Efficient heart disease prediction system using decision tree,in International Conference on Computing, Communication & Automation, 15-16 May 2015, Greater Noida, India
[18] Kabiraj, S.,2020. Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), July 1-3, 2020, Kharagpur, India.
[19] Dahiwade, D., Patle, G. & Meshram, E., 2019. Designing Disease Prediction Model Using Machine Learning Approach. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), March 27-29, 2019, Erode, India.
[20] Ambekar, S. , Phalnikar,R. , 2019, Disease Risk Prediction by Using Convolutional Neural Network,IN 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 16-18 Aug. 2018, Pune, India



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY