IJMRT: Volume (6), Issue 6, 2024
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN (Online):2663-4007

Review on Different Machine Learning Algorithms for Disease Prediction

Surbhi Agrawal[1], Pranavh Vummidi[2], Ravi Teja Devu[3], Rithin Ponduri[4*]

[1]*Associate Professor, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[2]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[3]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*
[4]*Student, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.*

[*]Corresponding author

DoI: https://doi.org/10.5281/zenodo.11316222

Abstract

This work aims to address the increasing complexity of healthcare diagnostics by developing a robust predictive model capable of accurately predicting diagnoses based on patient healthcare data. Leveraging diverse datasets encompassing various patient conditions and their interrelationships, the research seeks to create a comprehensive predictive system that assists medical practitioners in making timely and informed decisions. The Research involves constructing a machine learning model that learns from existing patient data to identify patterns and relationships between different health indicators and diagnoses. By analysing these patterns, the model can predict potential diagnoses for new patients, aiding healthcare professionals in diagnosis and treatment planning. In addition to model development, the research includes the creation of a user-friendly front-end application. This user-friendly interface enhances accessibility and usability, facilitating seamless integration into clinical workflows and promoting efficient decision-making in healthcare settings.

Keywords: Machine Learning, SVM Classifier, Naïve Bayes Classifier, Decision Tree Algorithm.

1. Introduction

In the modern era of healthcare, the ability to predict diseases based on specific symptoms plays a crucial role in early detection, timely intervention, and improved patient outcomes. This review focuses on developing a disease prediction system that utilizes machine learning techniques to analyze and predict diseases based on symptoms provided by the user. The goal is to create an automated and user-friendly application that empowers individuals to assess their health status and potential risks from the comfort of their homes.

By inputting their symptoms into the application, users can receive personalized predictions regarding possible diseases or health conditions they may be facing. Leveraging machine learning algorithms such as Decision Tree, Naive Bayes, and SVM, this application aims to harness the power of data-driven predictive modeling. These algorithms are trained on datasets containing information about symptoms and corresponding diseases, enabling them to learn complex patterns and relationships between symptoms and diseases.

Through the integration of Decision Tree, Naive Bayes, and SVM classifiers, this application offers users multiple approaches to disease prediction, enhancing the robustness and accuracy of the predictions. By providing insights into the likelihood of specific diseases based on input symptoms with the presence of one hundred and thirty-three symptoms used in dataset of final prediction out of forty-one unique diseases can be made, this application empowers users to take proactive measures for their health and seek appropriate medical attention if necessary.

With the increasing accessibility of technology and the growing reliance on digital solutions for healthcare management, this disease prediction application serves as a valuable tool in empowering individuals to make informed decisions about their health and well-being. By bridging the gap between symptom recognition and disease prediction, this application contributes to early detection, proactive healthcare management, and ultimately, improved

quality of life for users.

In conclusion, this disease prediction system leverages advanced machine learning techniques to provide users with accurate, personalized health insights based on their symptoms, facilitating early detection and proactive healthcare management.

1.1 Problem Statement

This research aims to provide accurate and timely diagnoses for patients by leveraging machine learning algorithms to analyze specified symptoms. By inputting symptoms into the system, users receive personalized predictions regarding potential diseases or health conditions. Through robust predictive modeling using algorithms such as Support Vector Machine, Naive Bayes, and Decision Tree, the system offers comprehensive analyses and conclusive diagnoses. This enables patients to make informed decisions about their health and seek appropriate medical attention promptly, ultimately leading to improved healthcare outcomes and better quality of life.

2. Literature Review

Importance of data preprocessing in disease prediction, emphasizing techniques like data cleaning, normalization, and feature extraction to improve model accuracy. Data cleaning involves removing or correcting errors and inconsistencies in the dataset, ensuring that the information is accurate and reliable. Normalization adjusts the data to a standard scale, which is crucial for algorithms that are sensitive to the scale of the input features.[1]

The use of domain-specific knowledge also extends to the identification of potential interactions between different symptoms and medical test results. By incorporating expert knowledge into the feature engineering process, the models can be tailored to capture nuances

in the data that generic algorithms might overlook. This careful and informed feature selection and creation process significantly contributes to the robustness and accuracy of the predictive models, as it ensures that the most relevant information is utilized in the disease prediction task.[2]

SVM algorithm has been widely adopted for its robustness in handling high-dimensional data. Support Vector Machine (SVM) is a powerful and versatile machine learning algorithm that has been widely adopted for its robustness in handling high-dimensional data [3]. One of the key strengths of SVM is its ability to utilize different kernel functions, such as linear, polynomial, and radial basis function (RBF) kernels. These kernels allow SVM to transform non-linearly separable data into a higher-dimensional space where a linear separation is possible.

Deep learning excels at finding complex patterns in large datasets, which could be particularly useful for disease prediction. By incorporating traditional machine learning methods alongside deep learning, investigation can be done about whether such a combination could lead to more robust and accurate disease prediction models.[4]

There are many machine learning algorithms (such as KNN, Random Forest and Decision Tree Classifier algorithms and many more) which were selected and on the given data many algorithms were applied so as to produce the best results. This research paper, will try to implement functions of machine learning in health facilities in a particular system [5].
Author analyzes data mining techniques which can be used for predicting different types of diseases. The reviewed research papers concentrate on predicting heart disease, Diabetes and Breast cancer etc., using various mining approaches [6].

In [7], the algorithms which are tested are J48 algorithm, Logistic model tree algorithm and Random Forest algorithm. This research compared different algorithms of Decision Tree classification seeking better efficiency in performance with respect to heart disease diagnosis using WEKA. The goal of this study was to identify hidden patterns by applying different mining techniques.

3. Proposed Methodology

Figure 1 data set is downloaded which has disease names and its symptoms associated with it. When designing the algorithm, we assumed that the client would have a clear understanding of the signs he was seeing. The constructed prediction takes into account 95 manifestations, and the customer might accept the signs of his preparation as input throughout that time.

	Disease	Count of Disease Occurrence	Symptom
0	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0008031_pain chest
1	NaN	NaN	UMLS:C0392680_shortness of breath
2	NaN	NaN	UMLS:C0012833_dizziness
3	NaN	NaN	UMLS:C0004093_asthenia
4	NaN	NaN	UMLS:C0085639_fall
...	...	...	...
1861	NaN	NaN	UMLS:C0425251_bedridden^UMLS:C0741453_bedridden
1862	NaN	NaN	UMLS:C0242453_prostatism
1863	UMLS:C0011127_decubitus ulcer	42.0	UMLS:C0232257_systolic murmur
1864	NaN	NaN	UMLS:C0871754_frail
1865	NaN	NaN	UMLS:C0015967_fever

[1866 rows x 3 columns]

Figure.1. Input Dataset

Figure 2 Pre-processing methods are applied to make nan data as usable data. Data pre-processing: The methods used in the mining of data that transform the raw data or re-encrypt it to create a structure so that it can be successfully decoded using computation are referred to

as information pre-processing. The following list includes the information pre-processing techniques used in the work that was just introduced:

A. Data Purification: This involves taking certain actions, such as adding back value that has been lost, to eliminate inconsistencies in the information.

B. Data Reduction: When dealing with a large information base, the investigation becomes challenging. So, we exclude those independent variables (symptoms) that might not have an impact on the objective variables (diseases). Which of the approximately 95 of 132 adverse effects that are clearly associated with the illnesses will be picked for the ongoing task

C. Models: The entire system is built to predict diseases using three algorithms, namely the Decision Tree model and the SVM classifier model. This allows the predictive analysis study to be proposed at the end of the study by examining the speed, efficiency, and performance of the different algorithms for the input dataset.

	disease	symptom	occurence_count
0	hypertensive disease	shortness of breath	3363.0
1	hypertensive disease	dizziness	3363.0
2	hypertensive disease	asthenia	3363.0
3	hypertensive disease	fall	3363.0
4	hypertensive disease	syncope	3363.0

Figure.2. Pre-Process Input Dataset

3.1. Split data into training and test data for training model and use test data to check the model

Input regarding an object is mapped to the item's output using the decision tree learning technique. Classification trees are tree models that have finite output classes. In these tree-like structures, the leaves represent class labels, while the branches depict the relationships between system class names and attribute values. Regression trees are decision trees with continuous output classes. A decision tree can be a decision-making input in data mining. Finally, from the recorded advancement of (ML) Machine Learning technique and the approaches in clinical area, it very well may be shown that frameworks and systems have been arisen that has empowered refined information investigation by basic and direct utilization of Machine Learning (ML) models. This paper brings an extensive near investigation of three models' execution of a clinical record with each of the obtaining accuracy score up to 98 %. Finally, the paper is investigated with disarray lattice & precision value.

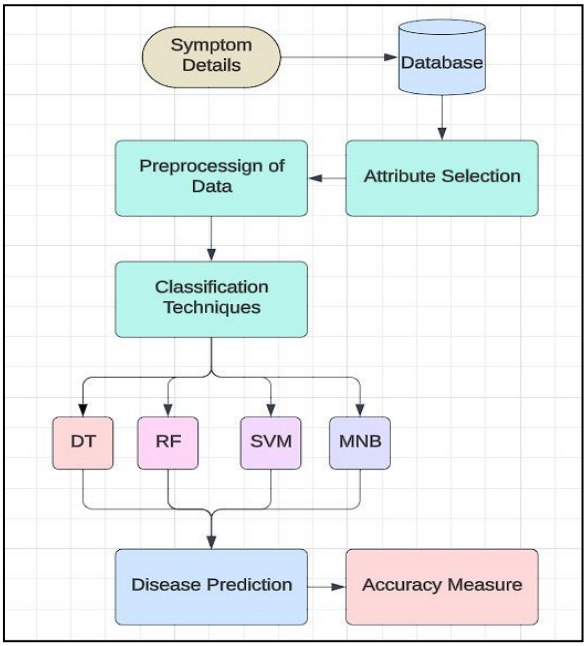

Figure.3. Workflow of the Research

4. Results and Discussion

When a framework is established with a prepared dataset and standardized datasets are generated using validated calculations, and client symptoms are inputted into the algorithm, and the symptoms are matched with the standardized dataset, this leads to the formulation of plans and the prediction of the most likely infection.

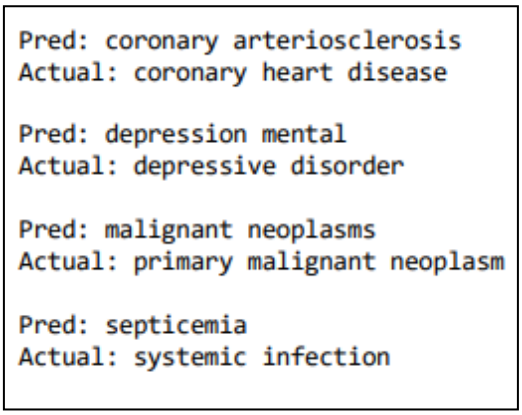

Figure.4. Testing model with test data

----------PERFORMANCE ANALYSIS FOR DT CLASSIFIER----------

	precision	recall	f1-score	support
HIV	1.00	1.00	1.00	1
bipolar disorder	1.00	1.00	1.00	1
cellulitis	1.00	1.00	1.00	1
cirrhosis	1.00	1.00	1.00	1
colitis	1.00	1.00	1.00	1
confusion	1.00	1.00	1.00	1
delirium	1.00	1.00	1.00	1
delusion	1.00	1.00	1.00	1
dementia	1.00	1.00	1.00	1
endocarditis	1.00	1.00	1.00	1
gastroesophageal reflux disease	1.00	1.00	1.00	1
glaucoma	1.00	1.00	1.00	1
hemiparesis	1.00	1.00	1.00	1
hepatitis	1.00	1.00	1.00	1
hiv infections	1.00	1.00	1.00	1
hypertensive disease	1.00	1.00	1.00	1
insufficiency renal	1.00	1.00	1.00	1
ischemia	1.00	1.00	1.00	1
kidney disease	1.00	1.00	1.00	1
kidney failure acute	1.00	1.00	1.00	1
lymphoma	1.00	1.00	1.00	1
malignant neoplasms	0.00	0.00	0.00	0
malignant neoplasms	1.00	1.00	1.00	1
manic disorder	1.00	1.00	1.00	1
melanoma	1.00	1.00	1.00	1
paranoia	1.00	1.00	1.00	1
paroxysmal dyspnea	1.00	1.00	1.00	1
pneumonia	1.00	1.00	1.00	1
primary malignant neoplasm	0.00	0.00	0.00	1
sepsis (invertebrate)	1.00	1.00	1.00	1
transient ischemic attack	1.00	1.00	1.00	1
accuracy			0.97	30
macro avg	0.94	0.94	0.94	30
weighted avg	0.97	0.97	0.97	30

Figure.5. Analysis for Decision Tree Classifier

The framework aims to forecast chronic diseases within a particular locality and demographic, utilizing the Decision Tree Algorithm along with Machine Learning techniques like Support Vector Machine, and Naive Bayes. With a notable accuracy rate of 93.4% for specific ailments.

	precision	recall	f1-score	support
HIV	1.00	1.00	1.00	1
bipolar disorder	1.00	1.00	1.00	1
cellulitis	1.00	1.00	1.00	1
cirrhosis	1.00	1.00	1.00	1
colitis	1.00	1.00	1.00	1
confusion	1.00	1.00	1.00	1
delirium	1.00	1.00	1.00	1
delusion	1.00	1.00	1.00	1
dementia	1.00	1.00	1.00	1
endocarditis	1.00	1.00	1.00	1
gastroesophageal reflux disease	1.00	1.00	1.00	1
glaucoma	1.00	1.00	1.00	1
hemiparesis	1.00	1.00	1.00	1
hepatitis	1.00	1.00	1.00	1
hiv infections	1.00	1.00	1.00	1
hypertensive disease	1.00	1.00	1.00	1
insufficiency renal	1.00	1.00	1.00	1
ischemia	1.00	1.00	1.00	1
kidney disease	1.00	1.00	1.00	1
kidney failure acute	1.00	1.00	1.00	1
lymphoma	1.00	1.00	1.00	1
malignant◆neoplasms	1.00	1.00	1.00	1
manic disorder	1.00	1.00	1.00	1
melanoma	1.00	1.00	1.00	1
paranoia	1.00	1.00	1.00	1
paroxysmal◆dyspnea	1.00	1.00	1.00	1
pneumonia	1.00	1.00	1.00	1
primary malignant neoplasm	1.00	1.00	1.00	1
sepsis (invertebrate)	1.00	1.00	1.00	1
transient ischemic attack	1.00	1.00	1.00	1
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

Figure.6. Analysis for SVM Classifier

	precision	recall	f1-score	support
HIV	1.00	1.00	1.00	1
bipolar disorder	1.00	1.00	1.00	1
cellulitis	1.00	1.00	1.00	1
cirrhosis	1.00	1.00	1.00	1
colitis	1.00	1.00	1.00	1
confusion	1.00	1.00	1.00	1
delirium	1.00	1.00	1.00	1
delusion	1.00	1.00	1.00	1
dementia	1.00	1.00	1.00	1
endocarditis	1.00	1.00	1.00	1
gastroesophageal reflux disease	1.00	1.00	1.00	1
glaucoma	1.00	1.00	1.00	1
hemiparesis	1.00	1.00	1.00	1
hepatitis	1.00	1.00	1.00	1
hiv infections	1.00	1.00	1.00	1
hypertensive disease	1.00	1.00	1.00	1
insufficiency renal	1.00	1.00	1.00	1
ischemia	1.00	1.00	1.00	1
kidney disease	1.00	1.00	1.00	1
kidney failure acute	1.00	1.00	1.00	1
lymphoma	1.00	1.00	1.00	1
malignant neoplasms	0.00	0.00	0.00	0
malignant◆neoplasms	1.00	1.00	1.00	1
manic disorder	1.00	1.00	1.00	1
melanoma	1.00	1.00	1.00	1
paranoia	1.00	1.00	1.00	1
paroxysmal◆dyspnea	1.00	1.00	1.00	1
pneumonia	1.00	1.00	1.00	1
primary malignant neoplasm	0.00	0.00	0.00	1
sepsis (invertebrate)	1.00	1.00	1.00	1
transient ischemic attack	1.00	1.00	1.00	1
accuracy			0.97	30
macro avg	0.94	0.94	0.94	30
weighted avg	0.97	0.97	0.97	30

Figure.7. Analysis for Naïve Bayes classifier

5. Future Scope

The aim of this forthcoming implementation is to develop a web-based platform designed for predicting disease outcomes using a variety of symptoms and conditions. Users will be able to select various symptoms and access disease predictions along with their probability data sourced from a comprehensive collection of datasets.

6. Conclusion

We aimed to create a user-input system to analyze symptoms, thereby easing strain on hospital OPDs and reducing the workload on medical staff. We successfully developed the system, integrating the above mentioned algorithms to achieve higher accuracy, averaging between 92% to 94%. Notably, we included a database feature for storing user data, which will aid in future enhancements. The system features an intuitive interface with visually appealing data representations. This review highlights the system's significance in transforming disease prediction methods, emphasizing its user-friendly design and potential for ongoing improvements.

REFERENCES

[1]. S. B. Kotsiantis and D. Kanellopoulos, Discretization Techniques: A Recent Survey, Artificial Intelligence Review, vol. 52, no. 1, pp. 665-715, 2019.

[2]. T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, and L. Cavedon, “The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges,” ACM Computing Surveys, vol. 55, no. 2, Art. 33, pp. 1-40, Feb. 2023.

[3]. M. H. Almaspoor, A. Safaei, A. Salajegheh, and B. Minaei, “Support Vector Machines in Big Data Classification: A Systematic Literature Review,” June 2021.

[4]. A. Esteva, A. Robicquet, B. Ramsundar, et al., “A Guide to Deep Learning in Healthcare,” Nature Medicine, vol. 25, pp. 24-29, 2019.

[5]. A. Rajkomar, J. Dean, and I. S. Kohane, “Machine Learning in Medicine,” The New England Journal of Medicine, vol. 380, pp. 1347-1358, 2019. DOI:10.1056/NEJMra1814259

[6]. M. Abdar, S. R. Niakan Kalhori, T. Sutikno, I. Subroto, and G. Arji, “Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases,” International Journal of Electrical and Computer Engineering (IJECE), vol. 5, pp. 1569-1576, 2015.

[7]. M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," Healthcare Analytics, vol. 3, 2023.