IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

MULTI DISEASE PREDICTION SYSTEM USING MULTI CRITERIAN DECISION MAKING

DHARANI.P
MSc (Decision and Computing Sciences) – Vth year
Coimbatore institute of technology
Coimbatore, India

S.GOWSIKKAN
MSc (Decision and Computing Sciences) - Vth year
Coimbatore institute of technology
Coimbatore, India

Dr.V.SAVITHRI
Assistant Professor
Dept. of Computing (DCS)
Coimbatore institute of technology
Coimbatore, India

ABSTRACT

In recent years machine learning has brought about a significant transformation in the field of healthcare. These advanced technologies have paved the way for more accurate and timely disease prediction and diagnosis, offering new hope in the battle against some of the world's most prevalent and life-threatening medical conditions. The realms of medicine and healthcare have witnessed a revolution driven by data-driven approaches. In an era where diseases like liver disorders, heart disease, breast cancer, and diabetes mellitus continue to affect millions of lives globally. These technologies empower healthcare professionals to not only predict disease occurrences with greater precision ultimately improving patient outcomes and reducing mortality rates. The escalating prevalence of liver disorders has raised concerns worldwide, making them a leading cause of mortality in numerous nations[1]. Heart disease, including cardiovascular ailments, remains a pressing global health issue. Researchers have turned to machine learning methods to predict heart diseases, striving to equip medical experts with the tools needed for early detection[3]. Breast cancer stands as one of the most prevalent and life-altering forms of cancer, predominantly affecting women worldwide[2]. Diabetes Mellitus is a chronic disease influenced by a myriad of factors, from genetics to lifestyle choices[4]. The research endeavors, offering detailed insights into the methodologies, results, and implications of harnessing machine learning techniques in the healthcare domain. These studies collectively underscore the transformative potential of advanced computational approaches in revolutionizing medical diagnosis and patient care, ultimately offering hope in the battle against these critical diseases.

Keywords: Machine Learning, decision-making.

I. INTRODUCTION

In the realm of healthcare, the significance machine learning in disease diagnosis and prediction cannot be overstated. Liver disorders have been on the rise, becoming a leading cause of mortality in several countries. This thesis embarks on a comprehensive investigation of liver patient datasets, employing machine learning techniques across three crucial

phases. In a similar vein, the paper explores diverse machine learning methodologies for heart disease prediction, addressing the pressing need to detect cardiovascular diseases

pre-emptively. Furthermore, breast cancer prediction and diagnosis take centre stage as the study applies five machine learning. In the context of diabetes, machine learning proves instrumental in enhancing prediction accuracy by incorporating external factors. Lastly, chronic kidney disease (CKD) is examined, utilizing predictive modelling to identify the most relevant attributes and employing 12 different machine learning-based classifiers[5]. This research underscores the pivotal role of data-driven approaches in advancing disease diagnosis and prediction, offering a promising avenue for improving healthcare outcomes.

II. LITERATURE SURVEY

M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi focuses on predicting liver disease using machine learning algorithms. The authors likely discuss the dataset, features, and the specific machine learning algorithms used for liver disease prediction. The primary emphasis is on evaluating the performance of these algorithms for liver disease diagnosis.[1]

Mohammed Amine Naji ,Sanaa El Filalib Kawtar Aarikac , EL Habib Benlahmard, Rachida Ait Abdelouhahide , Olivier Debauchef have centred around the application of machine learning for breast cancer prediction and diagnosis. It may discuss the dataset, features, and various machine learning models used for breast cancer detection.
The paper likely explores the accuracy and effectiveness of machine learning in diagnosing breast cancer.[2]

M. Preethi, Dr. J. Selvakumar have provide an overview of various studies and techniques related to heart disease prediction.[3]

Aishwarya Mujumdara , Dr. Vaidehi V predicts diabetes using machine learning algorithms. It may discuss the dataset, features, and the machine learning models used for diabetes prediction. The primary goal is to assess the efficiency of machine learning in diabetes prediction.[4]

Md. Ariful Islam,Md. Ziaul Hasan Majumder and Md. Alomgeer Husseinc predicting chronic kidney disease using machine learning algorithms. It may delve into the dataset, features, and the specific machine learning models employed for chronic kidney disease prediction. The emphasis likely lies on the performance and accuracy of these algorithms.[5]

G. Purusothama and P. Krishnakumari have done survey of data mining techniques used for predicting heart disease. It probably covers a broad range of data mining methods and their applications in heart disease risk prediction.[6]

B. Nithya and Dr. V. Ilango have focuses on predictive analytics in healthcare using machine learning.It may discuss various machine learning tools and techniques applied to healthcare, possibly covering a range of medical conditions.[7]

III. METHODOLOGY

The methodology employed for implementing the predictor model encompasses the following key stages: Data collection, Data preprocessing and Model validation.

3.1 Data Collection:

Data collection for various medical conditions has been conducted using datasets from Kaggle, providing valuable insights into disease prediction and diagnosis. For chronic kidney disease, the dataset includes parameters such as age, blood pressure, serum creatinine, and haemoglobin levels, enabling the development of predictive models to identify CKD cases. In the case of liver disease, patient data encompass factors like bilirubin levels and liver enzyme markers, serving as crucial indicators for hepatic health. Additionally, heart disease prediction is facilitated by data containing age, cholesterol levels, and electrocardiogram results, contributing to early detection efforts. Furthermore, diabetes prediction models utilize information such as glucose levels, blood pressure, and BMI, offering a comprehensive

approach to diabetes risk assessment. Lastly, breast cancer datasets incorporate various features, including tumor characteristics and cell attributes, fostering research in breast cancer diagnosis and prognosis. These diverse datasets from Kaggle serve as valuable resources for the development of machine learning and data mining models, ultimately enhancing the field of medical diagnosis and prediction.

3.2 Data Preprocessing and Analysis:

Data preprocessing involves cleaning and transforming raw data to make it suitable for analysis. This crucial step includes handling missing values, scaling features, and encoding categorical variables to ensure the data is ready for machine learning algorithms. Effective preprocessing enhances the accuracy and effectiveness of predictive models.

3.2.1 Data Cleaning:

Handling missing values is crucial in data preprocessing. Imputing missing data with mean, median, or mode is a common approach, preserving dataset integrity. However, removal of rows or columns with insignificant missing data may be necessary to avoid skewing results. The choice depends on the extent of missing data and the potential impact on analysis, ensuring reliable and accurate insights in data-driven tasks.

The kidney dataset has some missing values, the row with missing value is removed.

3.2.2 Data Transformation:

One-Hot Encoding: Convert categorical variables into binary vectors (0 or 1) using one-hot encoding to make them compatible with machine learning algorithms.

For Breast cancer and kidney disease the target column is transformed.

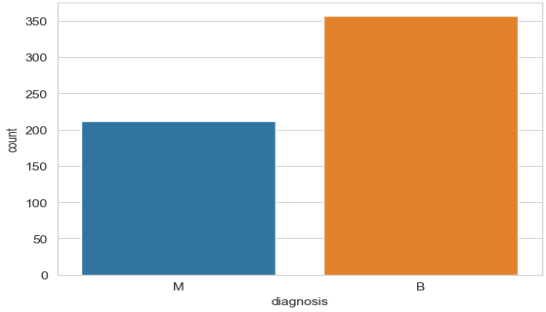

Fig 3.1 Breast cancer target bar chart

3.2.3 Data Splitting:

Splitting a dataset into training and testing sets is a fundamental step in machine learning and data analysis. This process allows us to evaluate the performance of our models and ensure that they can generalize well to unseen data. This partitioning ensures that the model learns from a substantial portion of the data while reserving a separate portion to assess its performance. By evaluating the model on the testing data, to determine its accuracy, precision, recall, or any other relevant metrics to understand how well it generalizes to new, unseen data. This process is critical for preventing overfitting and building models that can make reliable predictions. Divide the dataset into training and testing sets. Training data consists of 80% and test data consist of 20% for all the dataset.

3.2.4 Correlation analysis:

Correlation analysis is a statistical technique used to measure the strength and direction of the relationship between two or more numerical variables in a dataset. It quantifies how changes in one variable correspond to changes in another, aiding in the identification of associations or dependencies. Correlation coefficients, such as Pearson's correlation, range from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. This analysis is invaluable for feature selection, identifying redundant information, and gaining insights into data patterns, essential in data analysis, and machine learning model building. correlation analysis for the breast cancer dataset, the variables 'id,' 'symmetry_se,' 'smoothness_se,' 'texture_se,' and 'fractal_dimension_mean' were found to have weak or no significant correlations with other relevant features. Consequently, they were removed from the analysis, as they were deemed to contribute less valuable information and could potentially reduce noise and improve the performance of predictive models in breast cancer research

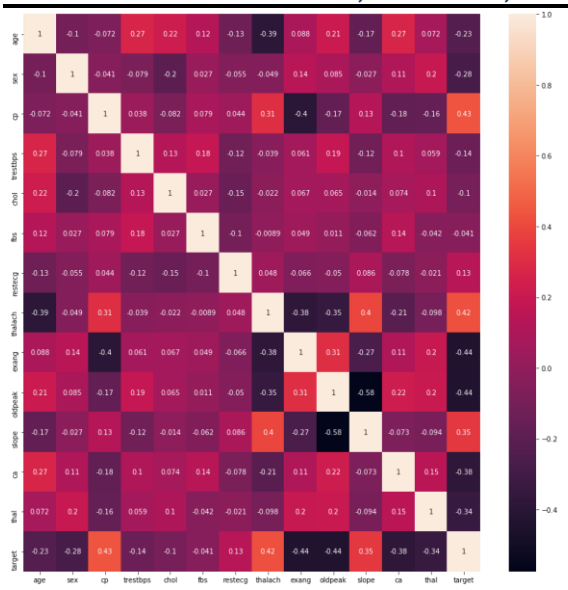

Fig 3.3 Heart dataset heatmap

3.3 Model Validation:

3.3.1 Logistic Regression:

Logistic regression, a widely-used classification algorithm in the context of medical disease prediction, leverages patient data features such as age, blood pressure, cholesterol levels, and various clinical indicators to determine the likelihood of a patient having a particular disease. By modelling the relationship between these input variables and the binary outcome (disease presence or absence), logistic regression estimates the probability of disease occurrence[]. It serves as an interpretable and efficient tool for medical practitioners to make informed decisions, aiding in early disease detection, risk assessment, and treatment planning based on a patient's individual health profile.

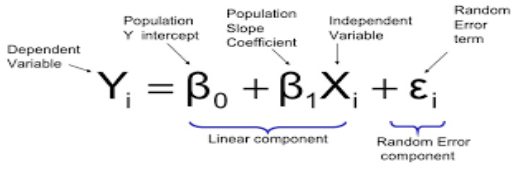

Fig 3.2 Formula for Logistic Regression

3.3.2 KNN:

K-Nearest neighbours (KNN) is a machine learning algorithm that can be effectively applied to medical datasets for disease prediction and diagnosis. In the context of medical data, KNN operates by measuring the similarity between a new patient's data and the data of existing patients within the dataset. It classifies the new patient based on the majority class among its K-nearest neighbors, where K is a user-defined parameter. This algorithm is particularly useful for medical applications as it considers the proximity of similar cases, making it well-suited for identifying patterns in diseases like chronic kidney disease, liver disease, heart disease, diabetes, and breast cancer. By leveraging KNN, healthcare practitioners can enhance the accuracy of disease prediction and improve patient outcomes, offering a valuable tool in the realm of medical data analysis.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

3.4 Formula for KNN

3.3.3 Random Forest:

Random Forest is a robust and versatile machine learning ensemble method extensively employed in medical disease prediction and diagnosis tasks. It operates by constructing a multitude of decision trees during training, each on a subset of the data and features. Through aggregating the predictions of these trees, Random Forest minimizes overfitting, enhances model generalization, and effectively handles diverse medical datasets with complex and interrelated features. In the context of liver disorders, heart disease, diabetes, breast cancer, and chronic kidney disease, Random Forest's ability to handle both categorical and numerical data, identify essential features, and produce reliable predictions has made it a valuable tool for healthcare practitioners and researchers, contributing to more accurate and interpretable disease assessments and ultimately improving patient outcomes.

3.3.4 Support Vector Machine:

Support Vector Machine (SVM) is a powerful machine learning algorithm employed in medical disease prediction and diagnosis. In the context of chronic kidney disease, liver disease, heart disease, diabetes, and breast cancer datasets, SVM stands out as a robust classifier capable of effectively separating patients into distinct classes based on various input features. SVM works by finding the hyperplane that maximizes the margin between different classes while minimizing classification errors. Its ability to handle both linear and non-linear data separation, coupled with its versatility in addressing binary and multi-class classification problems, makes SVM a valuable tool in medical data analysis. Through careful feature selection and preprocessing, SVM aids in the accurate identification of disease states, contributing significantly to early diagnosis and improved healthcare outcomes.

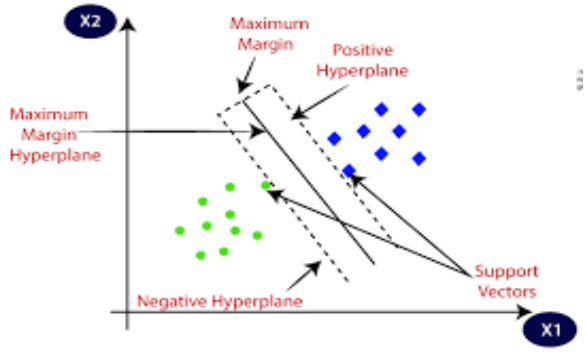

Fig 3.5 support vector machine

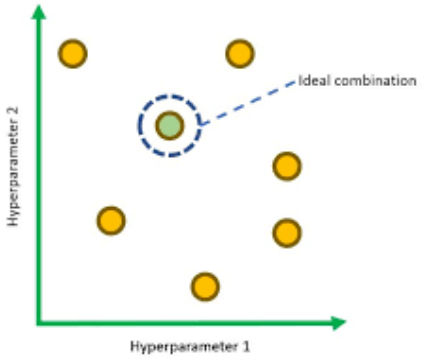

Fig 3.6 RandomizedsearchCV

3.3.5 Xgboost

XGBoost excels in this multi-disease context by handling diverse medical datasets with varying complexities. It optimally combines predictive features and copes with imbalanced data, ensuring accurate disease risk assessments. The algorithm's efficiency in handling high-dimensional and noisy medical data sets it apart, making it an ideal choice for identifying the risk factors, diagnosis, and prognosis of these life-altering conditions. With XGBoost, the system provides precise predictions, aiding healthcare professionals in early detection, personalized patient care, and tailored treatment strategies for diabetes, CKD, and heart disease."

3.3.6 RandomizedSearchCV

Hyperparameter tuning is a vital step in machine learning model optimization, enhancing model performance for tasks like heart disease classification. RandomizedSearchCV is an efficient approach that explores hyperparameter combinations randomly within specified ranges, reducing computational burden compared to grid search. It iteratively assesses a range of hyperparameters to find the optimal set, such as the learning rate, depth of decision trees, or regularization strength, resulting in improved accuracy, precision, and recall for heart disease and breast cancer prediction models.

3.4 Proposed System

The proposed system is an advanced medical diagnosis and prediction portal designed to revolutionize healthcare decision-making and patient care. Leveraging state-of-the-art machine learning techniques, this system is dedicated to providing accurate medical diagnoses, prognosis, and treatment recommendations**.**

A suite of machine learning models is employed to predict various medical conditions, including chronic diseases, infections, and rare disorders. These models are continually updated and refined to ensure accuracy.

The portal offers an intuitive user interface where patients and healthcare professionals can input patient-specific data. The system provides real-time predictions for medical conditions, helping both patients and healthcare providers make informed decisions regarding diagnosis and treatment plans.

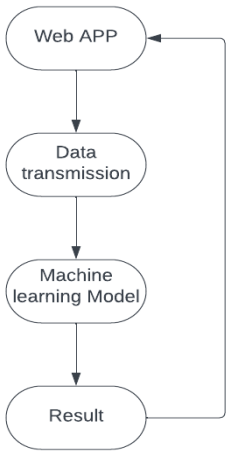

Fig 3.7 Flow chart of the system

Disease	Logistic Regression (%)	Random Forest (%)	K-Nearest neighbour's (KNN) (%)	Support Vector Classifier (SVC) (%)	Xgboost (%)	RandomizedsearchCV (%)
Liver Disease	71%	78%	83%	-		
Heart Disease	80%	95%				98%
Diabetes	76%	78%		80%	79%	
Breast Cancer	90%	91%	87%			96%
Chronic Kidney Disease	87%	99%				

Medical web application takes in medical data, the report data of patient information and symptoms. This data is then processed and transformed into a corresponding "disease pickle file" using machine learning models or algorithms. A "pickle file" typically refers to a serialized Python object, which contains the model's predictions or insights related to a specific disease.

The transformed data is served via a Flask web framework. Flask is a lightweight Python web framework used for creating web applications. It provides a means to serve the processed data as a web page or as an API endpoint, allowing users to access the results through a web interface. This process enables efficient and user-friendly access to medical insights and diagnoses based on the input data, facilitating better healthcare decision-making.

IV. ANALYSIS

V. IMPLEMENTATION

A comprehensive medical prediction portal has been developed using a combination of HTML, CSS, and Flask, offering an intuitive and powerful tool for healthcare predictions. This web application features a user-friendly input form where individuals can provide vital medical parameters, such as age, gender, symptoms, medical history, and diagnostic test results. Upon clicking the 'Submit' button, the entered data is seamlessly transmitted to a Flask application, which acts as the intermediary between the web page and a well-trained machine-learning model.
Within the Flask application, the input parameters are delivered to the machine learning model, specially designed for medical predictions. Leveraging advanced algorithms and vast medical datasets, this model accurately predicts various medical conditions, ranging from chronic diseases to infectious illnesses. The model then generates disease probability.

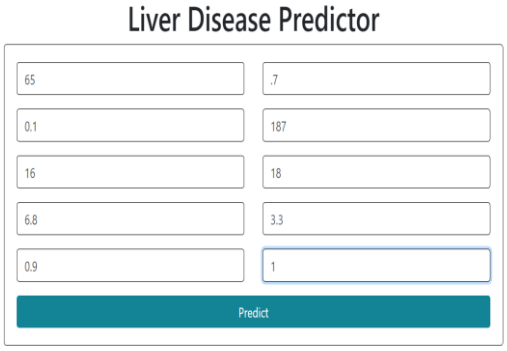

Fig 5.1 Form for liver disease

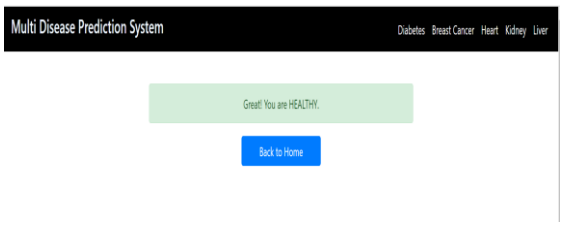

Fig 5.2 Result page for the liver data

VI. CONCLUSION

The integration of a multi-disease prediction system with a multi-criteria decision system in the field of medical diagnosis has shown remarkable promise. Across diverse datasets encompassing liver disease, heart disease, diabetes, breast cancer, and chronic kidney disease, machine learning models such as Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forest have demonstrated varying degrees of predictive accuracy. Among these models, Random Forest often excels by mitigating overfitting and handling complex, interrelated features. This approach empowers healthcare professionals and researchers with powerful tools for early disease detection and improved patient care. It enhances the potential for timely diagnoses, ensuring swift interventions, and ultimately raising the standard of healthcare. However, model suitability varies by disease and dataset characteristics, emphasizing the need for careful selection. In summary, the multi-disease prediction system, guided by multi-criteria decision systems, holds great promise for data-driven healthcare advancements, promising more accurate, interpretable, and reliable disease predictions that can significantly impact patient outcomes.

REFERENCE:

[1] M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms" International Research Journal of Engineering and Technology (2018).

[2] Mohammed Amine Naji ,Sanaa El Filalib Kawtar Aarikac , EL Habib Benlahmard, Rachida Ait Abdelouhahide , Olivier Debauchef "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis", International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT) , 2021

[3] M. Preethi, Dr. J. Selvakumar "A Literature Survey Of Predicting Heart Disease" International Research Journal of Engineering and Technology 2020

[4] Aishwarya Mujumdara , Dr. Vaidehi V "Diabetes Prediction using Machine Learning Algorithms" international conference on recent trends in advanced computing 2019.

[5] Md. Ariful Islam,Md. Ziaul Hasan Majumder and Md. Alomgeer Husseinc, "Chronic kidney disease prediction based on machine learning algorithms
" J Pathol Inform. 2023.

[6] G. Purusothama and P. Krishnakumari, "A Survey of
Data mining techniques on risk prediction: Heart disease", Indian Journal of Science and Technology, 2015.

[7] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems,2017.