IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107019 |

Polypathology Prediction Using Machine Learning

Sharvari Kadam[1], Adarsh M J[2]

PG Student, Dept. of Master of Computer Applications, Jawaharlal Nehru New College of Engineering, Shivamogga, India

Assistant Professor, Dept. of Master of Computer Applications, Jawaharlal Nehru New College of Engineering, Shivamogga, India

ABSTRACT: The burden of treating habitual conditions including diabetes, heart complaint, liver complaint, and cancer is mounting on healthcare systems each around the world. To ameliorate patient issues and save healthcare costs, numerous diseases must be detected and averted beforehand. Grounded on the data collected from cases, machine literacy algorithms have demonstrated considerable pledge for vaticinating the liability of contracting certain conditions. Data quality, point choice, and model interpretability are just a many of the issues and possibilities we cover in this area. We also draw attention to possible uses for these models in clinical settings. We conclude by relating new areas for disquisition and talking about the moral ramifications of using machine literacy to healthcare. A machine-based system for prognosticating multiple conditions has been developed using machine literacy and AI. The system uses techniques like point selection, pre-processing, dataset selection, and random forest algorithms. Results show that the proposed model outperforms other models in multi-disease vaticination, with Random Forest showing a 91.2 delicacy. This improves patient health and life expectancy.

KEYWORDS: Cardiovascular disease, Renal disease, Cancer, Diabetes, Hepatic disease, Machine learning,

I. INTRODUCTION

We frequently see individuals lose their lives as a result of not receiving timely care. Healthcare sectors are short on time and cannot decide which patients to treat first. However, healthcare sectors create massive amounts of data on people health. This data can provide a wealth of insights. So, we decided to create the project 'Multiple Disease Prediction System' utilising this data and modern machine learning algorithms. Our project is a collaboration of three machine learning models that will detect individuals with heart disease, renal disease, and diabetes at an early stage, allowing patients who are at risk to receive treatment first.

For our project, we first assessed the issue statement and identified what type of data would be needed. Kaggle provided us with three separate datasets for our three machine learning models. We collected data, analysed it, and visualised it for greater comprehension. The data was then cleaned by imputing null values and encoding category characteristics. On all datasets, we tested several classification techniques such as the Random Forest classifier and the XGBOOST classifier. We discovered that the Random Forest classification method performed better than the others. We obtained 98.52% testing accuracy on the heart disease dataset, 98.73% testing accuracy on the kidney diseases dataset, and 80.55% testing accuracy on the diabetes dataset using the Random Forest method. So we used Python's [Pickle library] to dump the model and [Python's Flask] framework to develop a web application for easy user interaction.

II. RELATED WORK

Here we have selected few key literatures after exhaustive literature survey and listed as below:

Sameer Meshram et al, [1] the paper presents a disease prediction system that utilizes the Naïve Bayes algorithm to predict the likelihood of various diseases based on a set of input features. The objective of the system is to assist healthcare professionals in making accurate and timely diagnoses.

Abid ishaq et al [2], the paper addresses the crucial task of predicting the survival rates of heart failure patients, which can aid healthcare professionals in making informed decisions regarding treatment and care.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107019 |

Bilal khan et al [3], the paper addresses the problem of predicting chronic kidney disease (CKD) using machine learning techniques. The authors aim to evaluate and compare the performance of different algorithms in order to identify the most effective approach for CKD prediction.

Chandrasekhar Rao Jetti et al [4], the paper presents a study on disease prediction using the Naïve Bayes machine learning algorithm. The authors aim to investigate the effectiveness of Naïve Bayes in predicting the occurrence of various diseases based on a given set of symptoms.

Selvaraj A et al [5], the paper present a prediction support system for multiple disease prediction utilizing the Naive Bayes classifier. The authors aim to develop a system that can assist in accurately predicting the presence or absence of various diseases based on patient symptoms.

Anjan Nikhil Repaka et al [6], the paper present a study on the design and implementation of a heart disease prediction system using the Naïve Bayesian algorithm. The authors aim to develop a reliable and accurate system that can predict the presence or absence of heart disease based on a set of input features.

Yashaswi G Sagar et al [7], the paper presents the MediInsight system, a smart health prediction system designed to assist in predicting health conditions and providing insights for better healthcare decision-making. The authors aim to develop a system that can leverage machine learning techniques to predict various health conditions accurately.

N. P. Tigga et al [8], the paper presents a study on the prediction of type 2 diabetes using machine learning classification methods. The authors aim to investigate the effectiveness of various machine learning algorithms in accurately predicting the presence or absence of type 2 diabetes based on patient data.

III. PROBLEM STATEMENT

Machine learning models for healthcare analysis often focus on a single disease at a time, causing issues with accuracy and time-consuming processes. However, multiple disease prediction can be used to anticipate multiple illnesses simultaneously, reducing the need for multiple websites. This approach is particularly useful for liver, diabetes, and heart disease, as they are all linked. By utilizing machine learning methods and Flask, numerous illness analyses can be created, allowing users to submit disease parameters and names, and Flask to execute the model and return the patient's status. This approach can significantly improve patient health and reduce the need for multiple websites for disease forecasting.

IV. DESIGN AND IMPLEMENTATION

The initial phase in system functioning is data gathering and selection of the most important properties. The required data is pre-processed into the required format. Following that, the data is divided into training and testing data. Machine learning techniques are applied, and training data is used to train the model. The system's correctness is determined by testing it with test data. This system is activated by utilising the modules listed below

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107019 |

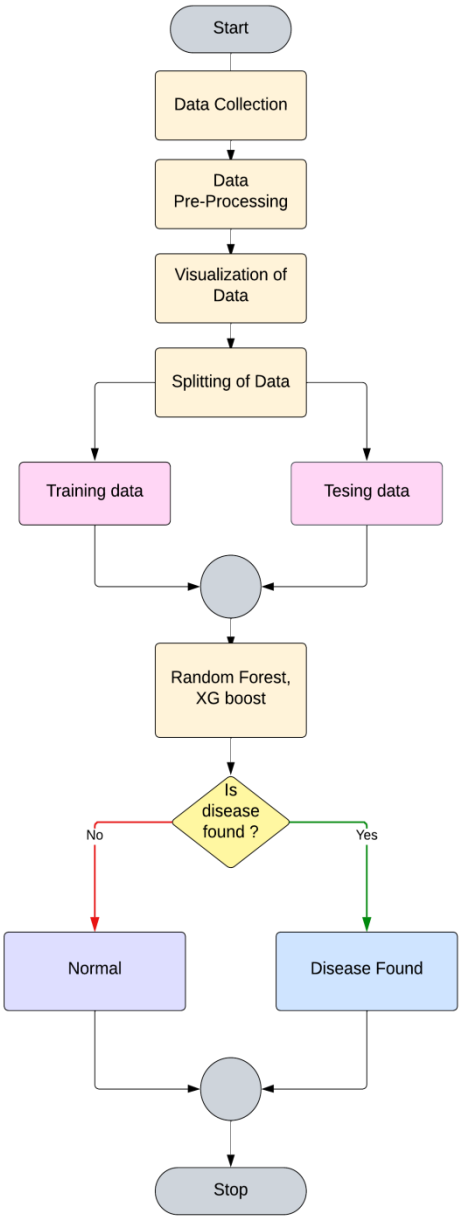

Figure 1: Flow chart of the system

The Fig.1 shows the flowchart of the multiple diseases detection using machine learning techniques. Data required for the prediction is collected using open resources.

Data collection: is an important step as the quality and quantity of the data that we gather for the proposed system will directly determine how good output that predictive model can be. The general approach is that we can collect data from open sources like Kaggle.

Data visualization: refers to the representation of data or information in a visual or graphical format. It involves creating visual images or graphics to convey complex data or information in a simple and easy-to-understand format. Visualization can be used to represent various types of data, including numerical, textual, spatial, and temporal data. The goal of visualization is to make data more accessible and understandable by representing it in a format that is easy to interpret and analyse.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107019 |

Splitting of data: It is the procedure for dividing the dataset into training and testing subsets, 75% of the data is used for training, while the remaining 25% is used for testing. Testing the data is used to evaluate the model's performance using the ML algorithm. The best model is chosen based on the training and testing data. Training data differs from testing data. The resulting data is sent into the Random Forest and Xg boost algorithms.

Model building & training: Supervised machine learning is one of the most successful and widely used forms of machine learning. When we need to predict a specific outcome or label from a set of supplied features and we have instances of feature-label pairs, we use supervised learning. These feature-label pairings, which comprise our training set, are used to build a machine learning model. Our goal for new, unheard-of data is to make precise forecasts. The terms "classification" and "regression" refer to the two major categories of supervised machine learning problems. Because detecting phishing urls is a continuous number, or in programming language, a floating-point number, our data set falls under the category of regression problem. Accuracy and F1 score are the metrics used to evaluate the model's performance.

Random Forest Algorithm: This machine learning algorithm is frequently used for disease prediction. It works by combining multiple decision trees to make predictions. For heart disease, it analyzes various features like age, blood pressure, and cholesterol levels to determine the likelihood of heart disease. Similarly, for cancer, it considers factors such as tumour size, cell type, and patient characteristics. In the case of kidney disease, it examines factors like creatinine levels and urine output. For liver disease, it analyzes variables such as bilirubin levels and liver enzymes. Lastly, for diabetes, it considers factors like glucose levels, BMI, and family history to predict the risk of developing diabetes. The algorithm uses these features collectively to generate accurate predictions for each respective disease.

XGB Classifier: XGBoost, an advanced machine learning algorithm, is used to predict liver disease by analysing various features such as liver enzymes, bilirubin levels, and other patient characteristics. It creates a boosted ensemble of decision trees, which collectively make predictions with high accuracy and can identify patterns and relationships in the data to determine the likelihood of liver disease.

Algorithm steps:

1. Select a random subset of samples from the given dataset.
2. Construct a Decision Tree using the selected subset of samples.
3. Repeat steps 1 and 2 multiple times to create multiple Decision Trees.
4. Predict new samples through Decision Trees for accurate predictions.
5. Aggregate Decision Tree predictions for final sample prediction.
6. Repeat steps 4 and 5 for each new sample to be predicted.
7. Assess Random Forest and XGB classifier model's performance using metrics like accuracy, precision, recall.
8. Optimize Random Forest and XGB classifier parameters, adjusting tree size and depth as needed.
9. Once the model is trained and validated, it can be used for disease prediction on new, unseen data.

These steps help create an ensemble of Decision Trees that work together to provide accurate predictions for disease classification.

V. RESULTS AND DISCUSSION

The application of Machine Learning in multiple disease prediction has shown promising results. By training models on diverse datasets containing patient information, symptoms, medical history, and test results, it becomes possible to accurately predict the presence or likelihood of different diseases. The predictive models developed through Machine Learning algorithms can achieve high accuracy rates, enabling healthcare practitioners to make informed decisions and tailor treatment plans to individual patients. The results of these models can significantly impact healthcare systems, as they streamline the diagnostic process, optimize resource allocation, and enhance patient care.

VI. CONCLUSION

Finally, the use of machine learning algorithms, specifically the Random Forest and Xgboost classifiers, for multiple disease prediction yielded promising results. Accurate predictions about the presence or absence of various diseases can be made using a dataset containing relevant medical features and training the classifiers on this data. Random Forest and Xgboost classifiers have both proven to be effective in dealing with complex datasets with multiple variables.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107019 |

These algorithms can capture non-linear relationships and interactions between features, which improves prediction accuracy. It has been discovered that these classifiers can effectively classify different diseases with high accuracy, sensitivity, and specificity after their implementation and evaluation. This suggests that they have the potential to be useful tools in medical diagnosis and disease detection.

REFERENCES

1. Sameer Meshram, Shital Dongre, Triveni Fole. "Disease Prediction System using naïve bayes". International Journal for Research in Applied Science & Engineering Technology Volume 10 Issue XII Dec 2022.
2. Abid ishaq, Saima sadiq, muhammad umer, saleem ullah, seyedali mirjalili, vaibhav rupapara , and michele nappi. "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques". March 16, 2021.
3. Bilal khan, Rashid naseem, Fazal muhammad, ghulam abbas, and sung hwan kim. "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy". March 30, 2020.
4. Chandrasekhar Rao Jetti, Rehamatulla Shaik, Sadhik Shaik, Sowmya Sanagapalli "Disease Prediction using Naïve Bayes - Machine Learning Algorithm", December 2021.
5. Prediction Support System for Multiple Disease Prediction Using Naive Bayes Classifier". Selvaraj A, Mithra MK, Keerthana S, Deepika M. International Journal of Engineering and Techniques - Volume 4 Issue 2, Mar-Apr 2021.
6. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin," Design and Implementing Heart Disease Prediction Using Naïve Bayesian", IEEE, June 2019.
7. Yashaswi G Sagar, Sahana Gajanana, Riyal Vivek, Swetha P," MediInsight: A Smart Health Prediction System", (IRJET), June 2021.
8. N. P. Tigga and S. Garg, ''Prediction of type 2 diabetes using machine learning classification methods,'' Proc. Computer Sci., Jan. 2020.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details