Multiple Disease Detection using Machine Learning

Jefrin Siby Mathew
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
jefrinsibymathew2023@cs.ajce.in

Joyal Joseph
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
joyaljoseph2023@cs.ajce.in

Roshik George
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
roshikgeorge2023@cs.ajce.in

Tinu Rose Thottungal
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
tinurosethottungal2023@cs.ajce.in

Honey Joseph
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
honeyjoseph@amaljyothi.ac.in

***Abstract*—The project "Multiple Disease Detection using Machine Learning" aims to develop a system for the accurate and efficient detection of multiple diseases using machine learning algorithms. The system is designed to analyze patient data, including medical history, symptoms, and test results, and predict the likelihood of several diseases simultaneously. The project involves data preprocessing, feature selection, and model training using various machine learning techniques such as decision trees, random forests, and support vector machines. The performance of the developed system is evaluated based on metrics such as accuracy, precision, recall, and F1-score using a dataset of patients with multiple diseases. The results of this project have the potential to improve the accuracy and efficiency of disease diagnosis, leading to better patient outcomes and reduced healthcare costs.**

***Index Terms*—Support Vector Machine, Logistic Regression, disease prediction, accuracy, precision.**

I. Introduction

The primary objective of a project centered around machine learning for the detection of multiple diseases is to create a system that can provide precise diagnoses by analyzing a variety of medical data inputs, including symptoms, medical history, and test results. This system leverages machine learning algorithms to evaluate and interpret the data to produce a diagnosis. The purpose of this project is to enhance the precision and efficiency of disease diagnosis, providing healthcare professionals with quick and reliable results. The project typically involves gathering a large amount of data, preprocessing it, training a machine learning model, evaluating its performance, and deploying it in a healthcare setting. The successful outcome of the project is dependent on several factors such as the quality of the data, the choice of algorithms, and the capability to handle the complexity of diagnosing multiple diseases. The impetus behind this project lies in the need to improve the accuracy and efficiency of disease diagnosis in the healthcare sector. Conventional diagnostic methods can be prone to errors and can be time-consuming, resulting in delayed treatment and increased costs. The aim of this project is to harness the potential of machine learning to enable quick and accurate diagnosis of multiple diseases by analyzing medical data inputs.

In the healthcare industry, data mining has become increasingly essential. When utilized effectively, data mining techniques can extract valuable information from large databases, aiding medical practitioners in making early decisions and improving health services. The classification can support physicians in the diagnosis process. Diseases such as malaria, dengue, Impetigo, Diabetes, Migraine, Jaundice, Chickenpox, and others can have a significant impact on an individual's health and may even lead to death if left untreated. To remedy this situation, the healthcare industry can leverage data mining algorithms such as support vector machine and logisitic regression to extract hidden patterns and relationships in their databases, allowing for effective decision-making. Therefore, we have developed an automated system that can extract and discover hidden knowledge associated with diseases from a historical (disease-symptom) database using the rule set of the respective algorithms.

II. Literature Survey

The article discusses the limitations of existing machine learning models for healthcare analysis, which focus on one disease per analysis, and proposes a system for predicting multiple diseases using Flask API. The system analyzes diabetes, diabetes retinopathy, heart disease, and breast cancer, and can be expanded to include other diseases. Machine learning algorithms, tensorflow, and Flask API are used to implement the system, and python pickling is used to save the model behavior. The system considers more parameters than existing models, which allows for more accurate detection of the effects of the disease. Flask API is designed to receive disease parameters and return the patient's status. The system aims to monitor patients and warn them in advance to decrease mortality rates. [1].

Heart disease is a prevalent and severe health concern that affects individuals across all age groups, and it can be caused by factors such as poor dietary habits, mental stress, smoking,

and more. Detecting heart issues at an early stage is crucial for receiving proper treatment, as treatment at a later stage can be expensive and risky. This paper focuses on utilizing machine learning techniques to predict heart disease using a set of health parameters collected from an individual. The study uses the heart disease dataset from the UCI machine learning repository and evaluates the heart disease prediction performance of four commonly used machine learning approaches, namely the naive Bayes classifier, KNN classifier, decision tree classifier, and random forest classifier. [2].

A Disease Prediction system that utilizes machine learning has been developed to forecast ailments based on the user's symptoms data, providing trustworthy findings. The system also suggests methods for maintaining good health and identifies sickness using the forecast. Due to a variety of diseases and a low doctor-patient ratio, the use of specific disease prediction technologies has become popular. The system aims to provide an instant and accurate disease prognosis based on the symptoms entered by the user, as well as the projected severity of the condition, using various machine learning algorithms to ensure speedy and reliable predictions. The system also provides the best algorithm and doctor consultation. [3].

Data mining techniques have been employed in various real-world applications such as healthcare, industry, and bioscience. The utilization of machine learning algorithms has enabled the extraction of useful information from medical databases, leading to the early prediction and diagnosis of diseases. In this research, a disease prediction system was developed using machine learning algorithms, including the Decision Tree classifier, Random Forest classifier, and Naïve Bayes classifier. The study used a sample data set of 4920 patients' records diagnosed with 41 diseases. The research work demonstrated the optimization of 95 independent variables (symptoms) closely related to diseases, and the comparative study of the results of the algorithms used. The aim of this research was to assist physicians in predicting and diagnosing diseases at an early stage. [4].

This systematic review explores the use of Machine Learning techniques in human disease diagnosis, focusing on modern techniques and their applicability in the medical field. The review aims to help researchers in various medical specialties to discover and choose the most appropriate algorithms for their specific situations. The authors conducted a thorough search of various databases and identified 141 relevant papers, which were analyzed and evaluated for their advantages and disadvantages. The review provides insights into the trends, goals, and areas of application of different Machine Learning techniques in medicine. [5].

Diabetes is a chronic disease affecting millions of people globally and leading to several health complications. Machine learning techniques is used to develop a system for early prediction of diabetes with higher accuracy. This paper involves combining results from different algorithms such as K nearest neighbor, Logistic Regression, Random forest, Support vector machine, and Decision tree to choose the best model for predicting diabetes. The accuracy of each algorithm is calculated to select the best one. [6].

Machine learning methods can be used to improve the diagnosis and assessment of Parkinson's disease (PD), as traditional diagnostic approaches may suffer from subjectivity and difficulty in classifying subtle motor symptoms. Additionally, early non-motor symptoms of PD are often overlooked, making early diagnosis challenging. Machine learning methods are used to classify PD and healthy controls or patients with similar clinical presentations. [7]

III. Methodology

The methodology for multiple disease detection using machine learning involves a combination of data preprocessing, feature selection, model selection, training, evaluation, tuning, and deployment. By following these steps, machine learning can provide an effective and efficient way to detect multiple diseases in patients.

A. Data Collection

The data collection phase of multiple disease detection using machine learning typically involves gathering and organizing relevant data from various sources, such as electronic medical records, medical imaging, laboratory test results, and patient self-reported data.

The following are the steps involved in the data collection phase:

- Define the scope and objectives of the project: The first step is to define the scope and objectives of the project, including the diseases to be detected, the type of data required, and the target population.
- Identify relevant data sources: The next step is to identify relevant data sources that contain the required data. This can include electronic medical records, medical imaging, laboratory test results, and patient self-reported data.
- Obtain necessary permissions and approvals: It is important to obtain necessary permissions and approvals from data owners, regulatory bodies, and ethical committees before collecting and using the data.
- Extract and preprocess the data: Once the necessary data sources have been identified, the data is extracted and preprocessed. This involves cleaning, formatting, and transforming the data into a format suitable for analysis.
- Label the data: In supervised learning, the data needs to be labeled with the corresponding disease or conditions. This is typically done manually by medical professionals.
- Store and manage the data: The data is stored and managed in a secure and accessible manner, following data privacy and security regulations.
- Validate the quality of the data: It is important to validate the quality of the data to ensure that it is accurate, complete, and relevant for the project.

Overall, the data collection phase is a critical step in multiple disease detection using machine learning, as it sets the foundation for the subsequent stages of data preprocessing, feature selection, and model training.

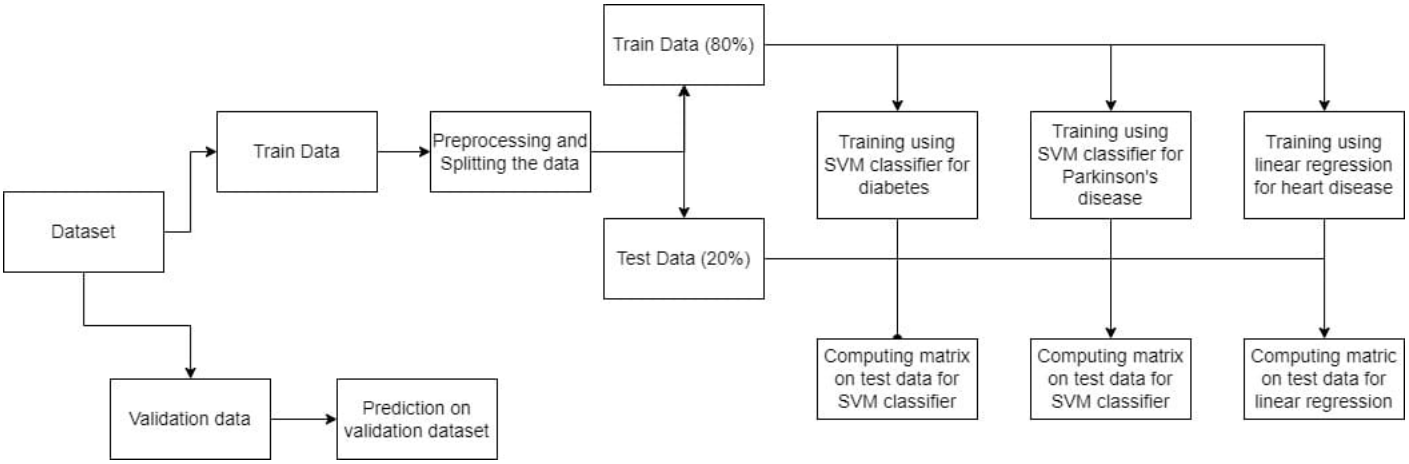

Fig. 1. Flow Diagram

B. Data Preprocessing

The data preprocessing phase is a crucial step in multiple disease detection using machine learning as it involves preparing the raw data for further analysis. The aim of this phase is to transform the data into a suitable format for model training and remove any inconsistencies or errors in the data.

The following are the steps involved in the data preprocessing phase:

- Data cleaning: This step involves removing any duplicate records, filling in missing values, and correcting any errors in the data.
- Data normalization: This step involves scaling and standardizing the data to ensure that all features are on the same scale. This can be achieved using techniques such as min-max scaling or z-score normalization.
- Data transformation: This step involves transforming the data to improve the performance of the machine learning model. This can include techniques such as log transformation, power transformation, or box-cox transformation.
- Feature selection: This step involves selecting the most relevant features that contribute the most to the prediction of the disease. Feature selection techniques can include statistical methods, correlation analysis, or machine learning algorithms.
- Data splitting: This step involves splitting the data into training, validation, and testing sets. The training set is used to train the machine learning model, the validation set is used to tune the hyperparameters of the model, and the testing set is used to evaluate the performance of the model.

Overall, the data preprocessing phase is critical to the success of multiple disease detection using machine learning. It ensures that the data is in a suitable format for analysis and improves the performance of the machine learning model by reducing noise and selecting the most relevant features.

C. Model Selection

The model selection phase in multiple disease detection using machine learning involves selecting the appropriate machine learning algorithm or model architecture to predict the presence of multiple diseases based on the preprocessed data. The choice of model is dependent on the characteristics of the dataset and the problem being solved.

The following are the steps involved in the model selection phase:

- Define the problem: Define the problem you want to solve using machine learning, including the target audience and the objective of the project.
- Select the appropriate type of machine learning: Choose the appropriate type of machine learning based on the problem you are trying to solve. For example, if you are predicting the presence of multiple diseases, you may choose a supervised learning approach.
- Choose a range of models: Select a range of machine learning models that are appropriate for the problem being solved. This can include techniques such as decision trees, random forests, support vector machines, neural networks, or ensemble models.
- Evaluate the models: Evaluate the performance of each model using appropriate evaluation metrics such as accuracy, precision, recall, or F1-score. This can be done using techniques such as cross-validation or hold-out validation.
- Compare the models: Compare the performance of each model and select the one that performs the best on the evaluation metrics. This can involve considering trade-offs between model complexity, interpretability, and performance.
- Tune the hyperparameters: Once a model has been selected, tune the hyperparameters to optimize the performance of the model. This involves selecting the values of the hyperparameters that result in the best performance on the evaluation metrics.

- Validate the model: Validate the performance of the selected model on a hold-out validation set to ensure that it generalizes well to new data.

Overall, the model selection phase is a critical step in multiple disease detection using machine learning as it determines the model's performance in predicting the presence of multiple diseases based on the preprocessed data.

1) Models Used:

- Support Vector Machine
  Support Vector Machine (SVM) is a type of supervised machine learning algorithm that can be used for classification or regression tasks. It is particularly useful when dealing with complex datasets that have multiple features. SVM works by creating a hyperplane that separates data into different classes. The basic idea of SVM is to find the hyperplane that maximizes the margin between the two classes. The margin is the distance between the hyperplane and the closest data points from each class. The hyperplane that maximizes this distance is the one that is most likely to generalize well to new data points. SVM works by mapping data into a high-dimensional space and then finding the hyperplane that separates the data into different classes. This is done by defining a kernel function that computes the dot product between two data points in the high-dimensional space. There are various types of kernel functions, including linear, polynomial, and radial basis function (RBF).SVM has several advantages, including its ability to handle high-dimensional datasets, its ability to find complex decision boundaries, and its ability to handle non-linearly separable data. However, it can be computationally expensive and may not work well on very large datasets.Overall, SVM is a powerful and widely used machine learning algorithm that can be used for a variety of classification and regression tasks.
- Logistic Regression
  Logistic Regression is a type of supervised machine learning algorithm that is commonly used for binary classification tasks, i.e., tasks where the output variable can take one of two values, such as yes/no, true/false, or 0/1. It is a variant of linear regression that uses a logistic function to model the relationship between the input variables and the output variable.The logistic function, also known as the sigmoid function, maps any input value to a value between 0 and 1. This makes it suitable for modeling probabilities, which is a key aspect of binary classification tasks. In logistic regression, the output of the model is the probability of the input belonging to the positive class (i.e., the class with a value of 1).To train a logistic regression model, the algorithm minimizes a cost function that measures the difference between the predicted probabilities and the actual labels in the training data. This is done using a gradient descent algorithm that adjusts the model's parameters (i.e., the weights and biases) iteratively until the cost function is minimized.Logistic regression has several advantages, including its simplicity, interpretability, and ability to handle noise and outliers in the data. It is also computationally efficient and can handle large datasets. However, it may not work well on datasets that have complex decision boundaries or require multi-class classification. Overall, logistic regression is a useful and widely used machine learning algorithm that is particularly well-suited for binary classification tasks where the goal is to predict the probability of an input belonging to the positive class.

D. Model Training and Testing

The model training and testing phase in multiple disease detection using machine learning involves training the selected model on the preprocessed data and evaluating its performance on a testing set. This phase aims to optimize the model's performance and ensure that it can accurately predict the presence of multiple diseases based on new data.

The following are the steps involved in the model training and testing phase:

- Split the data: Split the preprocessed data into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate the model's performance.
- Train the model: Train the selected model on the training set. This involves selecting appropriate hyperparameters and optimizing the model to achieve the best performance.
- Evaluate the model: Evaluate the performance of the trained model on the testing set using appropriate evaluation metrics such as accuracy, precision, recall, or F1-score. This step helps to determine the model's ability to generalize well to new data.
- Iterate: Iterate over steps 2 and 3 to refine the model further if necessary. This involves selecting different hyperparameters, changing the model architecture, or modifying the training process to improve performance.
- Validate the model: Validate the performance of the selected model on a hold-out validation set to ensure that it generalizes well to new data.
- Interpret the results: Interpret the results of the machine learning model and communicate the insights to the target audience. This can include visualizations, dashboards, or reports.

Overall, the model training and testing phase is crucial in multiple disease detection using machine learning as it optimizes the model's performance and ensures that it can accurately predict the presence of multiple diseases based on new data.

IV. Results

The results of multiple disease detection using machine learning refer to the performance of the trained model in predicting the presence of multiple diseases based on the preprocessed data. The results are typically evaluated using appropriate evaluation metrics such as accuracy, precision, recall, or F1-score.

The performance of the trained model depends on various factors such as the size and quality of the dataset, the chosen machine learning algorithm, the hyperparameters selected, and the training process used. If the model is effective, it should be able to accurately predict the presence of multiple diseases with high accuracy and precision.

The results of multiple disease detection using machine learning can be communicated through various means such as visualizations, dashboards, or reports. These can include a confusion matrix, receiver operating characteristic (ROC) curve, or precision-recall curve, which provide a visual representation of the model's performance.

Overall, the results of multiple disease detection using machine learning are crucial in assessing the model's effectiveness in predicting the presence of multiple diseases and providing insights into potential improvements to the model.

V. Conclusion

Multiple disease detection using machine learning is a complex and challenging project that involves various phases such as data collection, data preprocessing, model selection, model training and testing, and results and discussion.The project's main goal is to develop a machine learning model that can accurately predict the presence of multiple diseases based on a patient's medical history, symptoms, and other relevant information.The project's success depends on the availability and quality of the dataset, the selection of appropriate machine learning algorithms and hyperparameters, and the use of proper evaluation metrics to assess the model's performance.Multiple disease detection using machine learning has significant potential in improving the accuracy and speed of disease diagnosis, leading to better patient outcomes and reduced healthcare costs.Overall, this project is a promising area of research that has the potential to revolutionize the field of healthcare and provide significant benefits to patients and healthcare professionals alike.

References

[1] Akkem Yaganteeswarudu (2020). "Multi Disease Prediction Model by using Machine Learning and Flask API", the Fifth International Conference on Communication and Electronics Systems (ICCES 2020).

[2] B.Shiva Shanta Mani, V. M. Manikandan. "Heart Disease Prediction Using Machine Learning " in Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning (2021) pp. 373-381

[3] Kajal Patil, Sakshee Pawar, Pramita Sandhyan and Jyoti Kundale (2020). "Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques", ITM Web of Conferences 44, 03008 (2022) https://doi.org/10.1051/itmconf/20224403008 ICACC-2022.

[4] Sneha Grampurohit, Chetan Sagarnal(Jun 5-7, 2020). "Disease Prediction using Machine Learning Algorithms", 2020 International Conference for Emerging Technology (INCET) Belgaum, India.

[5] Nuria Caballé-Cervigón, José L. Castillo-Sequera, Juan A. Gómez-Pulido , José M. Gómez-Pulido and María L(26 July 2020). Polo-Luque."Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review", Received:31 May 2020; Accepted: 24 July 2020 by MDPI.

[6] KM Jyoti Rani(July-August-2020). "Diabetes Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science Engineering and Information Technology.

[7] Mohesh T, Gowtham K, Vijeesh P, Arun Kumar S (June 2022). "Parkinsons Disease Prediction Using Machine Learning", International Journal for Research in Applied Science Engineering Technology (IJRASET).