International Journal of Leading Research Publication
IJLRP

© IJLRP | ISSN (O) - 2582-8010
July 2024 | Vol. 4 Issue. 11
www.ijlrp.com

Predictive Modelling for Multiple Diseases Using Machine Learning

Kailash Yadav[1], Dr. Shantanu Bhattacharya[2]

[1]Student, [2]Associate Professor
Department of Information Technology,
KCC Institute of Technology and Management

ABSTRACT

This research project, titled " Predictive Modelling for Multiple Diseases Using Machine Learning " aims to predict several diseases, including diabetes, heart disease, lung cancer, Parkinson's disease, and breast cancer. The project employs various machine learning algorithms such as Support Vector Machines (SVM), logistic regression, random forest, and decision trees. For deployment, the models utilize Streamlit Cloud and the Streamlit library, which provide a user-friendly interface for disease prediction. The application interface offers five disease options: diabetes, Parkinson's disease, breast cancer, lung cancer, and heart disease. Users input relevant information for the selected disease, and the application swiftly generates a prediction result, indicating whether the individual may be affected by the condition. This research leverages machine learning techniques to meet the need for accurate disease prediction, facilitating early detection and preventive measures. The intuitive interface provided by the Streamlit library and Streamlit Cloud enhances accessibility and usability, allowing users to easily assess their risk for various diseases. The high accuracy of the different models demonstrates the effectiveness of the machine learning algorithms employed in disease prediction.

Keywords: Machine learning, deep learning, datasets, streamlit, heart disease prediction, diabetes detection, lung cancer detection, breast cancer detection.

Introduction

The Multiple Disease Prediction System using Machine Learning is an advanced healthcare tool designed to forecast the likelihood of multiple diseases based on patient symptoms. The project, titled "Multiple Disease Prediction using Machine Learning, Deep Learning, and Streamlit," aims to predict five specific diseases: diabetes, heart disease, Parkinson's disease, breast cancer, and lung cancer. To achieve this, the system employs various machine learning algorithms tailored to each disease: Support Vector Machines (SVM) for diabetes, Random Forest for Parkinson's disease and breast cancer, and Neural Networks for heart disease and lung cancer.

The application is deployed using Streamlit Cloud and the Streamlit library, providing a user-friendly interface. The process starts with collecting relevant data from Kaggle.com, which is then pre-processed for training and testing the prediction models. Each disease is predicted using the most appropriate algorithm for that condition. The application interface presents five disease options. Upon selecting a disease, users are prompted to input the required parameters for the corresponding model. The application then generates and displays the prediction result based on these inputs. Streamlit Cloud hosts and shares the application, making it readily accessible, while the Streamlit library facilitates the development of interactive and intuitive models.

Objectives:

- **Early Detection and Diagnosis:** To enable early identification of diseases.
- **Risk Assessment and Prevention:** To assess risk levels and implement preventive measures.
- **Personalized Medicine:** To tailor treatment plans to individual patient profiles.
- **Resource Optimization:** To make efficient use of healthcare resources.
- **Clinical Decision Support:** To assist healthcare providers in making informed decisions.

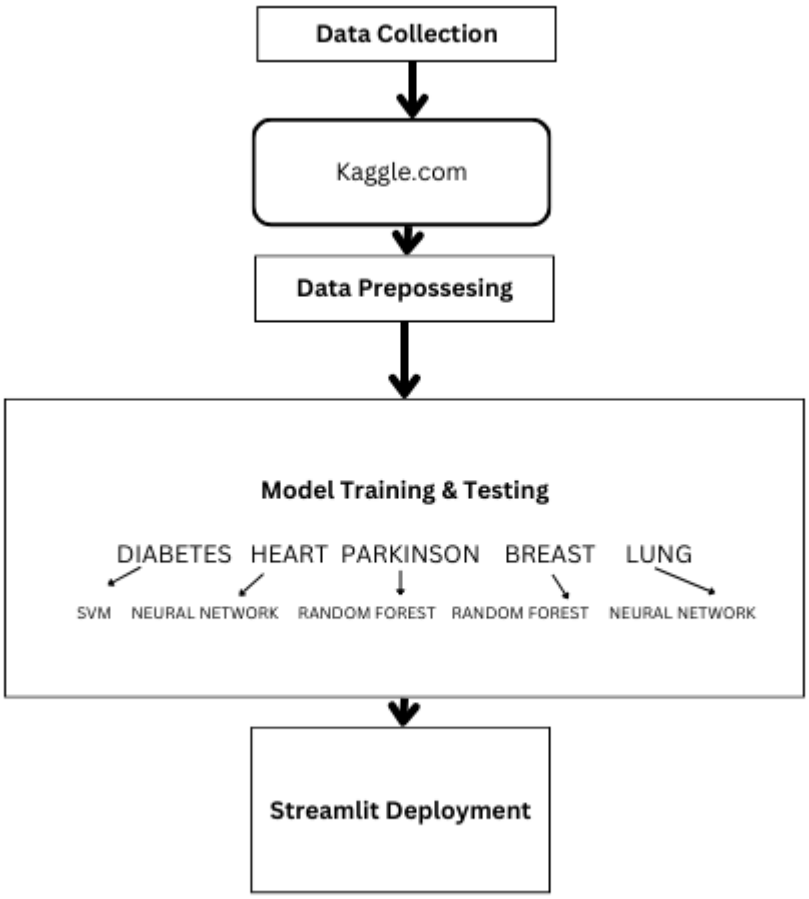
© IJLRP | ISSN (O) - 2582-8010
July 2024 | Vol. 4 Issue. 11
www.ijlrp.com

- **Public Health Surveillance:** To monitor and analyze disease patterns for public health insights.

Methodology

Here is a concise description of the methodology used in the Multiple Disease Prediction Project:

1. **Data Collection:** Data is sourced from Kaggle.com, which provides a variety of datasets. Specific datasets related to diabetes, heart disease, lung disease, Parkinson's disease, and breast cancer are collected.
2. **Data Preprocessing:** The collected data undergoes preprocessing to enhance quality and suitability for machine learning models. This involves managing missing values, removing duplicates, and normalizing or scaling features.
3. **Model Selection:** Different machine learning algorithms are chosen for each disease prediction task. Algorithms such as Support Vector Machines (SVM), Logistic Regression, Neural Networks, and Random Forest are selected based on their effectiveness for the respective diseases.
4. **Training and Testing:** The preprocessed data is divided into training and testing sets. Models are trained on the training data and evaluated on the testing data. Performance metrics are used to assess the accuracy and effectiveness of each model.
5. **Model Deployment:** An interactive web application is developed using Streamlit and deployed via Streamlit Cloud. The application features an intuitive interface with five disease prediction options—heart disease, lung cancer, diabetes, Parkinson's disease, and breast cancer. Users input relevant information to receive disease predictions based on their entries.



© IJLRP | ISSN (O) - 2582-8010
July 2024 | Vol. 4 Issue. 11
www.ijlrp.com

Proposed System

In our approach, we utilized various methods to enhance model performance and ensure effective disease prediction. These methods include:

- **Dimensionality Reduction:** This technique reduces the number of features in the dataset while retaining essential information. It helps in simplifying the model and improving computational efficiency.
- **Label Encoding:** This converts categorical textual data into numerical values, making it suitable for machine learning algorithms.
- **Data Standardization:** This ensures that the data is scaled consistently, which helps in improving model performance.

To predict multiple diseases—such as diabetes, heart disease, Parkinson's disease, breast cancer, and lung cancer—we employed several machine learning algorithms including Support Vector Machines (SVM), Logistic Regression, Neural Networks, Decision Trees, and Random Forest. Each algorithm was chosen based on its effectiveness for the specific disease prediction task.

The system is designed with a user-friendly interface using Streamlit, which facilitates easy interaction and input of parameters by users. The predictions are then generated based on the inputs provided. The application is deployed on Streamlit Cloud, making it accessible and efficient for users.

Data for the models was sourced from Kaggle, a renowned platform in the data science community. The data underwent thorough preprocessing to ensure quality and appropriateness for training the models. After preprocessing, the machine learning algorithms were trained and tested to evaluate their accuracy in predicting diseases.

Input & Output Design

Input Design:

The Multiple Disease Prediction System requires users to input specific data for each disease they wish to predict. The system is designed to be intuitive and user-friendly, guiding users to enter the necessary parameters based on their selected disease. Here is how the input design is structured:

1. **Sidebar Menu Options:**
    - **Diabetes**
    - **Parkinson's Disease**
    - **Breast Cancer**
    - **Lung Cancer**
    - **Heart Disease**
2. **Data Collection for Each Disease:**
    - When a user selects a disease from the sidebar menu, the application prompts them to enter relevant parameters for that specific condition. Each disease has a unique set of required inputs, which are designed to be significant for accurate predictions.
3. **Parameter Input:**
    - The application will display form fields or input controls tailored to the selected disease. For instance:
        - **Diabetes:** May require inputs such as blood glucose levels, BMI, age, and family history of diabetes.
        - **Parkinson's Disease:** Could ask for details on motor symptoms, tremor severity, and balance issues.
        - **Breast Cancer:** Might need information on age, family history, mammogram results, and symptoms.

© IJLRP | ISSN (O) - 2582-8010
July 2024 | Vol. 4 Issue. 11
www.ijlrp.com

- **Lung Cancer:** Could include factors such as smoking history, age, and symptoms like persistent cough.
- **Heart Disease:** May require inputs like blood pressure, cholesterol levels, age, and physical activity levels.

4. **User Interface:**
    - An easy-to-navigate interface will guide users to input their data. Each parameter will have a clear label, and validation checks will ensure that the data entered is in the correct format.

Output Design:

The output design of the Multiple Disease Prediction System aims to provide users with a clear and understandable result regarding their disease prediction. Here's how the output is structured:

1. **Result Presentation:**
    - After entering the necessary parameters and submitting them, the system will display a result indicating whether the user is likely to be impacted by the chosen disease.
2. **Output Format:**
    - The results will be presented in a straightforward and accessible format. For instance:
        - **Positive Result:** "Based on the provided data, you are at high risk for [Disease Name]. We recommend consulting with a healthcare professional for further evaluation."
        - **Negative Result:** "Based on the provided data, you are not currently at high risk for [Disease Name]. However, regular check-ups are advised."
        - **Further Information:** The system might also provide additional information or recommendations, such as lifestyle changes or follow-up tests, based on the result.
3. **Visualization:**
    - The results may include visual aids like charts or risk graphs to help users better understand their risk levels.
4. **Feedback:**
    - The system could offer options for users to provide feedback or request more detailed information about their results.

This design ensures that users receive accurate predictions in a user-friendly manner, allowing them to make informed decisions about their health.

Prediction result - Positive

"Prediction: The person is affected by [Disease Name]"

Prediction result - Negative

"Prediction: The person is not affected by [Disease Name]."

The output needs to be shown on the user interface so that the user can quickly understand the outcome of the prediction.
Overall, the output design clearly displays the prediction result on the user interface, and the input design makes sure the user can enter the required parameters for disease prediction.

© IJLRP | ISSN (O) - 2582-8010
July 2024 | Vol. 4 Issue. 11
www.ijlrp.com

Result

Here the result of applying different types of algorithms is shown below:

DISEASES	SVM	RANDOM FOREST	DECISION TREE	LOGISTIC REGRESSION	NEURAL NETWORK
DIABETES	78%	72%	64%	-	77%
HEART	-	82%	88%	83%	91%
PARKINSON	87%	94%	74%	-	89%
BREAST	-	96%	-	60%	94%
LUNG	-	-	-	94%	96%

Here's a summarized overview of the findings for each disease prediction based on the applied machine learning algorithms and their accuracies:

Diabetes Detection:

- **Algorithms Applied:** Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression, Neural Network
- **Highest Accuracy:** Support Vector Machine (SVM)

Heart Disease Prediction:

- **Algorithms Applied:** Random Forest, Decision Tree, Logistic Regression, Neural Network
- **Highest Accuracy:** Neural Network

Parkinson's Disease Prediction:

- **Algorithms Applied:** Support Vector Machine (SVM), Random Forest, Decision Tree, Neural Network
- **Highest Accuracy:** Random Forest

Breast Cancer Detection:

- **Algorithms Applied:** Random Forest, Logistic Regression, Neural Network
- **Highest Accuracy:** Random Forest

Lung Cancer Detection:

- **Algorithms Applied:** Logistic Regression, Neural Network
- **Highest Accuracy:** Neural Network

Summary:

- **Diabetes:** SVM performed the best in accuracy.
- **Heart Disease:** Neural Network achieved the highest accuracy.
- **Parkinson's Disease:** Random Forest was the most accurate.
- **Breast Cancer:** Random Forest had the highest accuracy.
- **Lung Cancer:** Neural Network provided the best results.

This detailed breakdown shows which algorithms performed best for each specific disease prediction, helping to understand the effectiveness of different machine learning techniques in healthcare applications.

© IJLRP | ISSN (O) - 2582-8010
July 2024 | Vol. 4 Issue. 11
www.ijlrp.com

Conclusion

This report presents the findings from various studies focused on enhancing patient-doctor interactions through advanced technology. Our proposed system is designed to address the gap between patients and healthcare providers, facilitating more effective communication and collaboration. By leveraging innovative technological solutions, our system aims to optimize the experience for both patients and doctors, ensuring that their objectives are met with greater efficiency and effectiveness.

1. **Machine Learning for Disease Prediction**
    - The system employs various machine learning methods to support the prediction of multiple diseases. This approach ensures accurate and timely diagnosis, benefiting patients by providing early detection and treatment options.
2. **Trust Development**
    - Unlike many current systems that merely automate processes, our system focuses on building trust with users. This is achieved through transparent and reliable predictions, making users more confident in the system's capabilities.
3. **Incorporating Physician Recommendations**
    - To further ensure user trust and maintain the integrity of doctors' practices, our system integrates physician recommendations. This dual approach ensures that while patients receive accurate predictions and advice, doctors' businesses are not adversely affected. Instead, it enhances their practice by providing a tool that supports their decision-making process.

By addressing these critical areas, our system aims to improve the overall healthcare experience for both patients and doctors, fostering a more collaborative and trustworthy environment.

References

1. Machine Learning for Healthcare: Review, Opportunities, and Challenges Authors: Chen M, Hao Y, Hwang K, Wang L, Wang L. Publication: IEEE Transactions on Cybernetics, 2017.
2. Title: Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records Authors: Miotto R, Li L, Kidd BA, Dudley JT. Publication: Scientific Reports, 2016.
3. Title: Predictive Modelling of Hospital Readmission Rates using Electronic Medical Record-wide Machine Learning: A Case-Study using Mount Sinai Heart Failure Cohort Authors: Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Publication: Pacific Symposium on Biocomputing, 2017.
4. https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8896926/
6. https://ieeexplore.ieee.org/document/9791739