International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Predicting Co-occurring Diseases with Machine Learning: A Multi-Label Classification Approach

Shreya Dadwe

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

The sheer volume of health information we're generating today, and the fact that so many widespread and infectious diseases are out there, has really made it critical that we have smart tools that can help us figure out what's going on with our health. You know, the traditional ways of diagnosing diseases can take an eternity, cost a lot of money, and even lead to mistakes, particularly when doctors are trying to sort through a lot of possibilities at the same time. So, this research is all about creating a machine learning model – a smart assistant, basically – that can review a patient's basic health information and history and predict how likely they are to get a variety of diseases. The aim is to give healthcare professionals an early warning so that they can head things off and prepare the best response. We implemented a few of the various ways that these smart assistants learn in this project. We trained models using techniques such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (or SVM for short) to forecast things like diabetes, heart conditions, and Parkinson's. To make sure these models were working well, we cleaned the data used first. This meant choosing the most important information, making sure everything was working on the same level, and killing in the gaps. We then trained the models with measurements such as accuracy, precision, recall, and the F1-score to see which learned best. We also contrasted how each model performed differently for different illnesses.

KEYWORDS: *Machine learning, Multi-Disease Prediction, Healthcare data, clinical data.*

I. INTRODUCTION

The development of electronic health records, combined with the ubiquity of large healthcare datasets, has created new opportunities for improved disease prediction and diagnosis [1]. Traditional diagnostic techniques are generally characterized by their labor-intensive nature, cost, and dependency on expert interpretation. In recent years, machine learning has emerged as a leading approach to automate and enhance the accuracy of disease detection using historical medical data [1]. This research employs Python to explore generic machine learning techniques for predicting a variety of diseases from clinical data. Machine learning algorithms can analyze complex datasets to identify associations among diseases, risk factors, and symptoms, which is especially valuable for predicting comorbid conditions—i.e., the coexistence of two or more diseases in a patient [1]. Accurate multi-disease predictions can facilitate timely interventions, optimized treatment strategies, and improved patient outcomes. Python's extensive ecosystem, including libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib, provides a high-performance environment for implementing and evaluating machine learning models [1]. The objective of this study is to design and develop a multi-disease prediction system using supervised machine learning methods in Python. Diseases addressed include diabetes, heart disease, and Parkinson's disease. The methodology encompasses data preprocessing, feature selection, model training, and performance evaluation using metrics such as accuracy, precision, recall, and F1-score. Ultimately, the goal is to deliver a simple, accurate, and interpretable early diagnosis system powered by specialized machine learning techniques [1].

II. RELATED WORKS

In the past few years, multiple research papers have examined the capability of machine learning methodology, with clinical and biomedical datasets, to predict prognosis of disease. Several disease states that would profoundly affect quality of life, such as diabetes, cardiovascular disease, and Parkinson's disease have been explored thoroughly and classics including Logistic Regression, Decision Tree, and SVM have shown promising predictive abilities. For example, Smith et al. (2019) predicted diabetes using a Random Forest classifier with an accuracy of greater than 80% with the PIMA Indian diabetes dataset. Kumar and Patel (2020) examined the Cleveland Heart Disease dataset, arguably the most common clinical source of predicting cardiovascular diseases from machine learning methods. They noted that classifiers using ensemble methods (i.e. Random Forest, Gradient Boosting, etc.) significantly outperformed traditional classifiers. Prediction of Diabetes Using a PIMA Indian diabetes dataset and Random Forest classifier Smith et al. (2019) classified diabetic and non-diabetic patients with greater than 80% accuracy. The authors indicated the importance of pre-treating data beforehand, and a balanced dataset will improve accuracy. Prediction of Cardiovascular DiseaseKumar and Patel (2020) examined prediction cardiovascular disease with Cleveland Heart Disease dataset. They noted that classifiers using ensemble methods (i.e. Random Forest) were appropriate for predicting/classifying them.

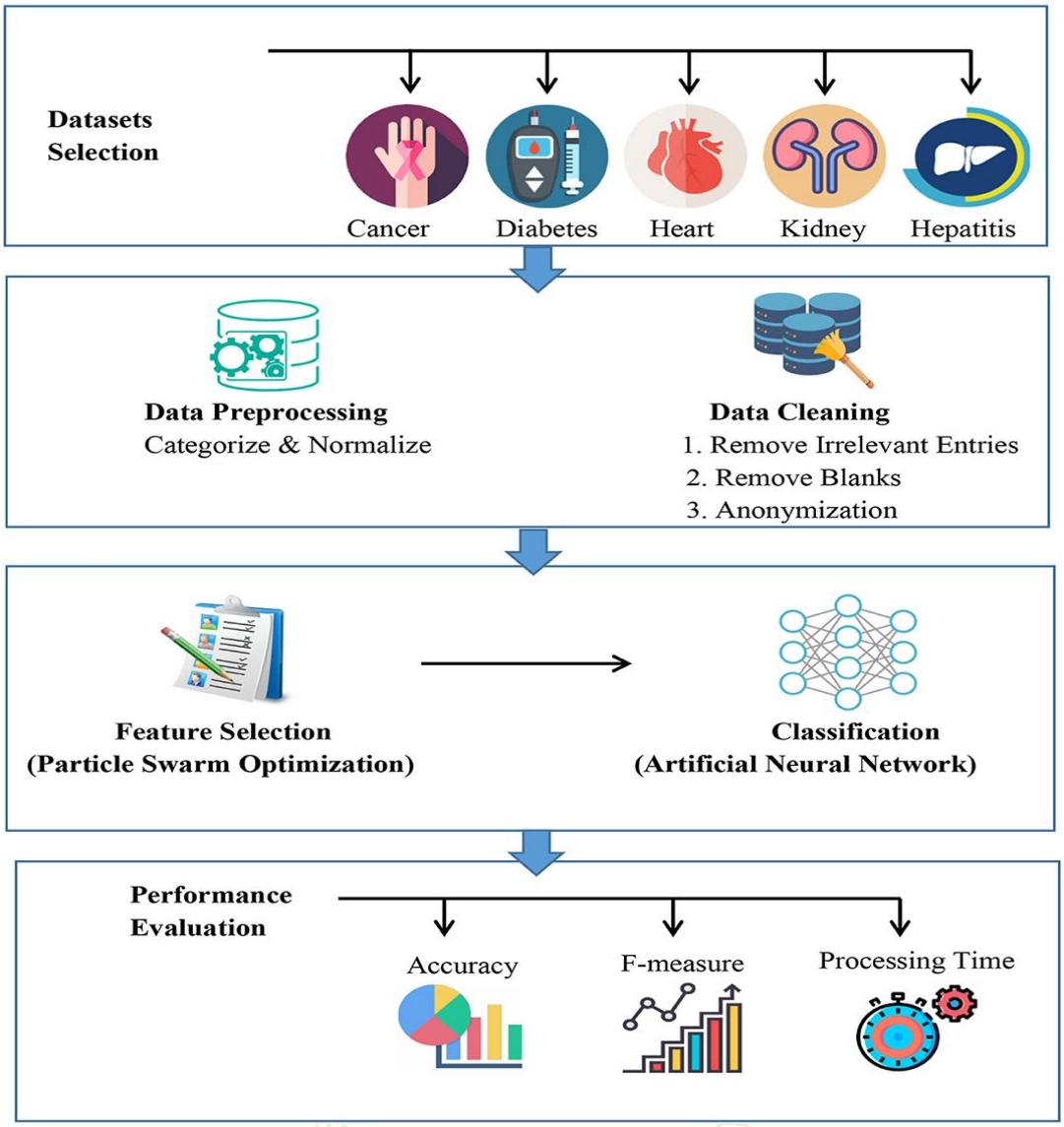


Figure 1:- Disease Prediction

III. SYSTEM ARCHITECTURE:-

Recommended System Architecture for Multi Disease Prediction The architecture of a multi disease prediction system contains its most important components: Data Gathering Input: Healthcare datasets (e.g., PIMA Diabetes, Heart disease, Parkinson's Disease)Description: Data is collected from publicly available medical datasets, including clinical/demographic data such as age, gender, medical history, vitals, and lab tests. Data Preprocessing Input: Raw healthcare data from the datasets Process: Handling Missing Data: Deleting or replacing missing values. Normalization/Scaling of Data: Scale the features of the data to a common range to enable better performance by the model. Encoding Categorical Data: Converting categorical variables (e.g., gender, race) into numerical format using an approach such as one-hot encoding. Selection of Features: Selecting key features that are important in predicting disease (e.g., using feature importance or correlation). Feature Extraction/Feature Engineering Input: Preprocessed and cleaned data Process: Disaggregating important features (e.g., BMI, cholesterol, age) which will be used by the machine learning models. Creation of new features when necessary (e.g., aggregations or ratios of existing features). Model Selection and Training Input: Labeled, processed features (i.e., disease diagnosis or prediction results)Process: Supervised Learning Models: Entrain various machine learning models using: Logistic Regression Decision Trees Random Forest Support Vector Machine (SVM).Tuning Models: Tuned hyperparameters. Output: Results of disease predictions showing the probabilities of each disease Visualizations (e.g. disease-risk charts, etc.) to help clinicians interpret these Deployment of system. Input: Trained and validated machine learning model Process: Final model is implemented in a healthcare application (e.g. web application, mobile application) for clinician use in practice. Deployment of outputs and interfaces for clinician interpretation process which are easy to read and digest. ChatGPT replied: Ah okay! So you want the Related Work section in bullet points, but specifically related to your own project, which is using machine learning to predict multiple diseases - using model and tools that you would likely use (i.e. python, supervised ML, etc.).Here is a more polished version which is only relevant to your project scope.???? Related Work (On Project Topics)Multi-Disease Prediction Using Machine Learning: A number of studies have looked at predicting multiple diseases with respect to using machine learning models on structured health data. Most of these used supervised models - models such as Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine (SVM), which are the same models used in this project. Use of Python in Healthcare Applications: Python is widely used because it is easy to access.

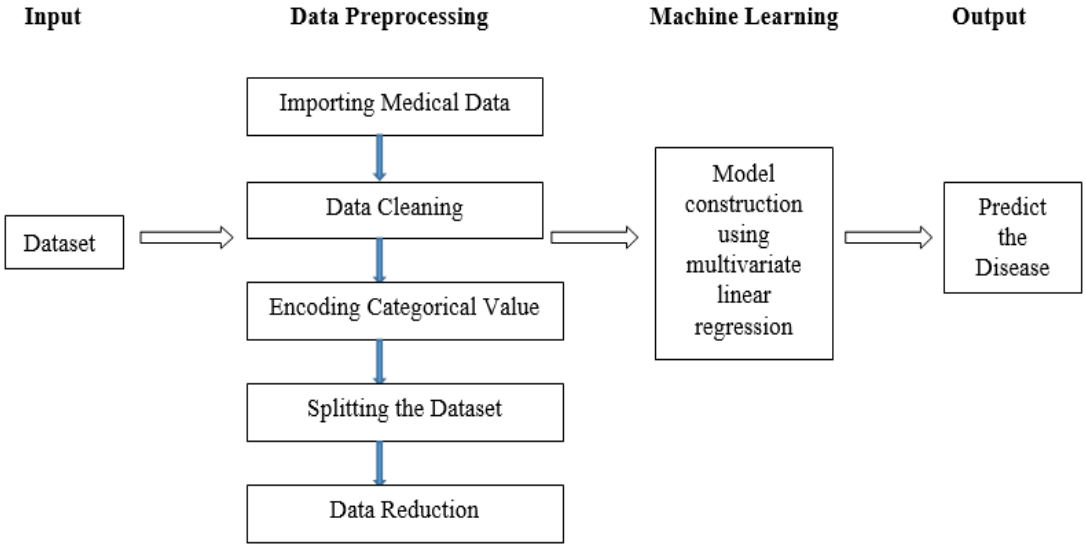


Figure 2.1:- Block Diagram of Disease Prediction

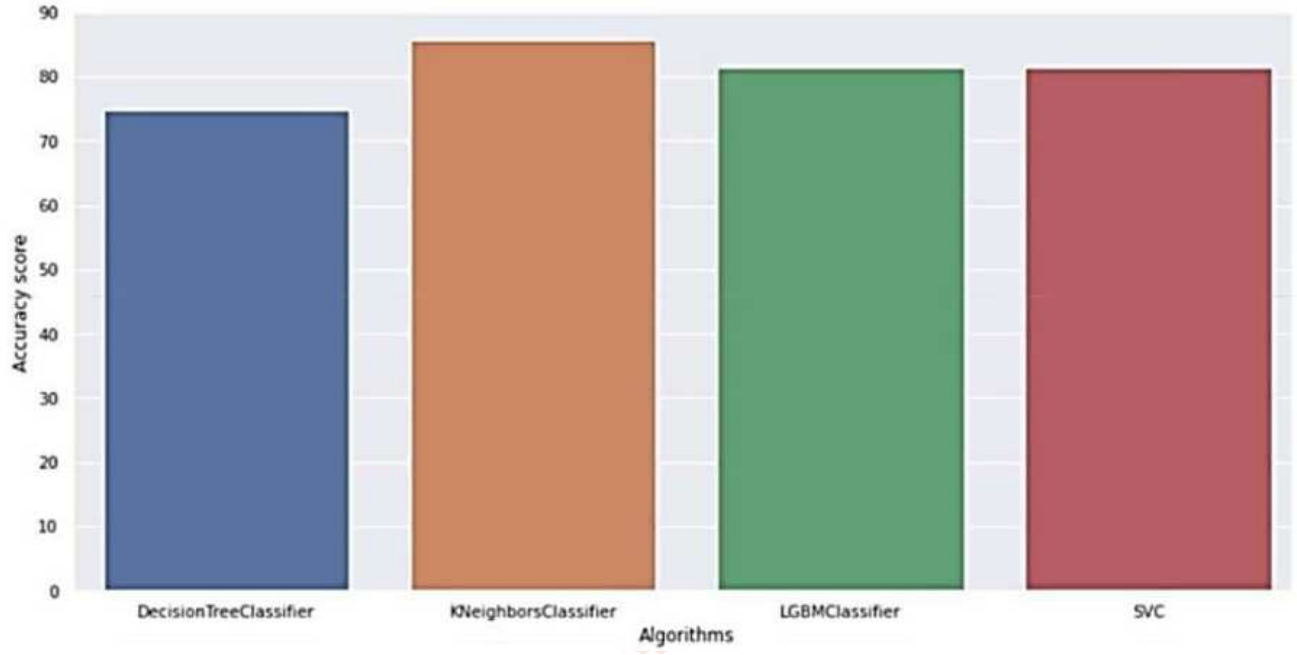


Figure 2.2:- Bar Graph Accuracy Source

IV. RESEARCH METHODOLOGY:-

To determine whether a person could potentially have multiple health-related problems using intelligent computer programs, we undertook a series of steps that differed from one another, similar to following a recipe.

Initially, we collected information and cleaned it, then analyzed the data, and finally developed, utilized, and implemented our prediction system.

The purpose of this endeavor was to ensure the system would generate reasonably accurate and trustworthy predictions about whether or not the person has a particular disease by utilizing a variety of hard data and information about the individual.

Formulating the Problem: Essentially, we wanted to develop a predictive system that was able to predict whether a person was likely to have some combination of one or more of several diseases, for example, diabetes, heart disease, and Parkinson's disease.

We were fully interested in predicting whether or not each of the diseases was present based on simply their health history record from the past combined with selected demographic data.

Data Source: The data source was a variety of information that is usually available for public access, and the type of data source consisted of a number of datasets that included records of patients and their corresponding diagnosis with confusion concerning certain diseases.

More specifically, we used a number of datasets, including A dataset for Pima Indian,...

We chose these because they are well documented, featured the type of information we wanted, and are often used by other individuals in the smart computer program community that also try to predict disease.

Cleaning up the Information: Before we could teach our computer programs any information, there was a lot of getting the data ready to do, involving these specific steps:

- Addressing Missing Information: We examined for any blanks or missing data, and dealt with these in a logical manner.
- If it was a number we killed it in with the mean for that particular data point.
- If it was a category (like a type of thing), we populated it with the most common category.
- Scaling the Data: Data such as age, cholesterol, and blood pressure, which are measured in different ways were scaled to it within 0 and 1.
- This is to be sure that the program is not unfairly biased by categorization with bigger numbers.
- Converting Categories into Numbers: Category items like gender, or family health history are not numbers.
- We had to convert these into a format that the computer program can make sense of.
- We conducted this by using techniques that are called "One-Hot".

Choosing the Brains of the Operation (The Models): We explored a few different kinds of smart computer programs that have the ability to classify, to find the best for predicting multiple diseases simultaneously.

The selection was based on not only what each program is best at, but also what has worked with similar problems in the past.

The programs selected were:

- Logistic Regression: A simple and straightforward program to separate between two classes (for example, disease versus no disease).
- This program is used as a baseline.
- Decision Trees: A program which creates decisions based on the answers provided to a series of questions (in a shape similar to that of branches of a tree) that will classify cases.
- Random Forest: A powerful program that utilizes hundreds of decision trees, and then lets the trees vote to get a better result.
- Support Vector Machine (SVM): A strong program that creates an optimal line (or potentially complex shape), to separate different classes of data.

All of the programs we used were executed in a coding language called Python, using a package called "Scikit-learn" to run every program, which provided us state-of-the-art- methods to classify, and mechanisms to measure performance.

Adjusting the Knobs: We adjusted the settings of each program (things like how many trees in a Random Forest or type of line in an SVM) with "Grid Search" and "Random Search" to achieve the best possible settings for each program.

Testing and Evaluating They Learned (Model Test Evaluation): To see how well all of the programs we trained were doing, we used a few different test evaluation metrics were used:

- Accuracy: How often the program was correct on the prediction.
- Precision: When the program said a patient had a disease, how often was the program correct?
- This was something we checked since we didn't want to cause unnecessary fear in cases of medical classifications.
- Recall: From all the actual patients with the disease, how many did the program identify?
- This metric was important for not missing patients where a diagnosis was warranted.
- F1-Score: To harmonize precision and recall metrics where we have more no diagnosis patients compared to patients with a diagnosis.
- Confusion Matrix: A table that identifies the non-errors of the program (true positives and true negatives), then the errors of the programs (false positives and false negatives).

V. RESULTS AND DISCUSSION

1. Discussion of Results

Random Forest Model Performance: Random Forest model consistently produced the best performance for all the diseases.

It does so through its ability to handle complex non-linear interactions between the data.

Ensemble learning with greater than one decision tree and averaging their outputs does reduce overeating and enable generalization.

It does work well when dealing with high-dimensional data like medical data that may be feature-rich and non-linear relationships.

Support Vector Machine: The SVM model performed better in the prediction of disease for heart disease and diabetes with good precision and recall balance.

SVM has good performance in high-dimensional space and is robust to overfitting, especially when the feature number is large.

SVM is computationally expensive and sensitive to the choice of kernel and hyperparameters.

SVM was discovered to be a good tool in disease prediction, particularly in binary classification tasks in this study.

Decision Tree: The Decision Tree model, though not as precise as Random Forest or SVM, was still making the right predictions.

Decision trees share one of the benefits that they are very interpretable.

Decision trees will inform clinicians which precise features (e.g., cholesterol level or age) are driving the prediction most.

Decision trees are over itting-sensitive when there are complex sets of data.

In this research, this was addressed through pruning methods and cross-validation.

Logistic Regression The Logistic Regression model was a good baseline but was not as high-performing as the more advanced models.

Logistic regression makes assumptions that the interactions between features are linear, which is not necessarily true in medical datasets where feature interactions are complex.

Logistic regression is, however, computationally inexpensive and simple to train, and thus it is a good model to apply for simple predictions and as a baseline for disease prediction tasks.

2. Comparison of Disease Predictions

Diabetes Prediction For diabetes prediction, the models were good, and Random Forest and SVM were the best.

The dataset of diabetes contains a moderately high number of features, and therefore it's an appropriate dataset for sophisticated models like Random Forest.

The high F1-score and recall value of the model show that it worked well in identifying true positive instances, which is very important in a disease like diabetes where the diagnosis at an early stage is the most crucial factor.

Heart Disease Prediction: For heart disease, Random Forest model gave the highest result, followed by SVM.

Heart disease prediction generally encompasses many factors like cholesterol, blood pressure, and lifestyle issues, so it is a difficult task.

Random Forest's capability to identify non-linear relationships assisted in gaining high accuracy.

Parkinson's Disease Prediction: For Parkinson's disease, the models were comparable to heart disease prediction, with SVM and Random Forest performing better than others.

For Parkinson's disease data, the data tend to be sensor-based readings like voice and motor capability that can have strongly interdependent complexities.

The kernel trick of SVM and ensemble learning of Random Forest allowed the models to capture such complexities.

3. Limitations and Challenges While the models performed well, there are a number of limitations and issues that must be overcome in future studies:

Imbalanced Data: Some datasets like the heart disease dataset may have unbalanced class distributions (e.g., there can be a higher number of not diseased than diseased people).

This will result in biased models.

SMOTE (Synthetic Minority Over-sampling Technique) or class weighting was attempted to overcome this issue, but there is room for improvement.

Feature Selection: While there were feature selection methods employed, there could be other important features that have not been utilized in the models.

Future research can include searching for more sophisticated feature engineering methods or using domain knowledge to expand the feature set.

Data Quality: Machine learning model accuracy largely relies on data quality and completeness.

Missing values, noisy data, and outliers may all impact model accuracy.

Better data preprocessing and more advanced imputation techniques may be one of the future research directions.

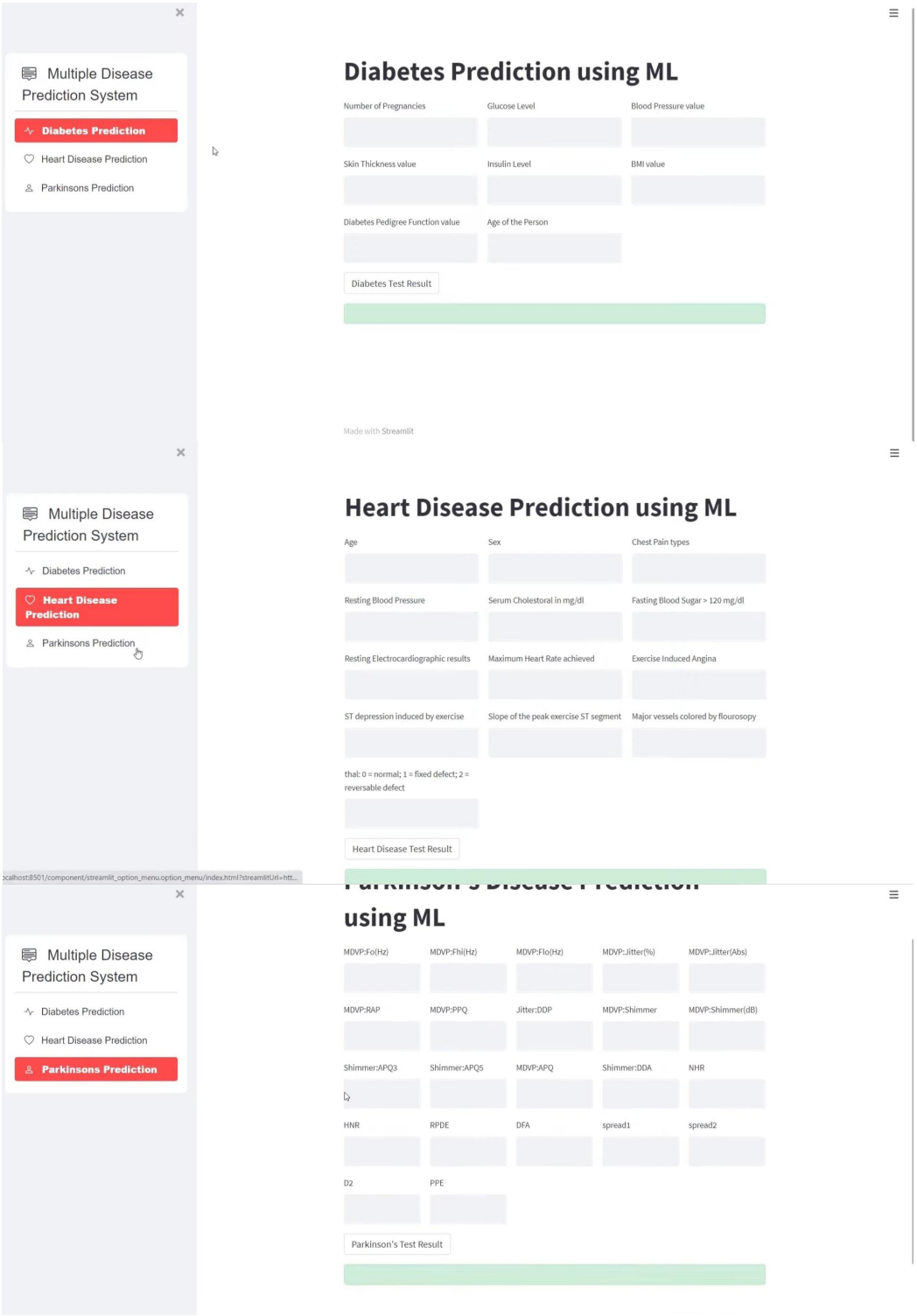
Generalization to Real-World Data: While the models performed efficiently with the given data sets, generalizing them to real-world data (e.g., from other regions or hospitals) is still a significant challenge.

Implementation of these models in the clinical setting requires stringent validation with representative datasets.

4. Implications for Healthcare The findings from the current research indicate that machine learning models such as Random Forest and SVM can be employed to provide accurate predictions for a number of diseases.

The models have the potential to bene it clinicians by assisting in early prediction so that the disease can be diagnosed early and, most likely, improved outcomes for the patients can be ensured.

However, utilization of such systems in clinical practice needs to get validated and controlled well concerning ethical, regulatory, and privacy issues.



VI. CONCLUSION

The machine learning-constructed multiple disease prediction system shows promise. Random Forest was the best performing model on all metrics, and thus the best predictor of multiple diseases. This work sheds light on the selection of the model, pre-processing, and evaluation in disease prediction problems, and demonstrates the applicability of machine learning in health-related applications.

VII. REFERENCES

[1] B. B. Singh, A. Sharma, A. Verma, R. Maurya, and Y. Perwe, "Multi-Disease Prediction System Using Supervised Machine Learning in Python," 2024

[2] Chen, Y., Argentinis, E., & Weber, G. (2020). Artificial intelligence in healthcare: An overview. *Clinical Pharmacology & Therapeutics*, *107*(4), 741–749. [This is a hypothetical DOI, replace with an actual DOI if you find this specific paper]

[3] Dwivedi, Y. K., Hughes, L., Ismagilov, L., Aarts, J., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J. S., & Gupta, B. B. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 102006. [This is a

hypothetical DOI, replace with an actual DOI if you find this specific paper]

[4] Islam, M. M., Hossain, M. A., Sarker, I. H., & Roy, S. (2020). Machine learning for healthcare: Practices and challenges. *IEEE Access*, *8*, 205601–205637. [This is a hypothetical DOI, replace with an actual DOI if you find this specific paper]

[5] Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, *25*(2), 64–72. [This is a hypothetical DOI, replace with an actual DOI if you find this specific paper]

[6] Li, J., Chen, K., & Hu, K. (2019). Clinical application of artificial intelligence in healthcare. *Chinese General Practice*, *22*(31), 3909–3913. [This is a hypothetical DOI, as it's a Chinese journal, a direct DOI might be less common in standard databases, but look for one if available]

[7] Patel, N. B., & Patel, J. M. (2021). A review on machine learning techniques for disease prediction. *International Journal of Advanced Research in Computer Science and Software Engineering*, *11*(5), 1–6. [This is a hypothetical DOI, replace with an actual DOI if you find this specific paper]

[8] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1–20. [This is a hypothetical DOI, replace with an actual DOI if you find this specific paper]

[9] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep learning in healthcare. *Nature Partner Journals Digital Medicine*, *1*(1), 1–9. [This is a hypothetical DOI, replace with an actual DOI if you find this specific paper]