Multiple Disease Prediction Using Machine Learning & Stremleat

1st Supriya. S. Nalawade
Department of CSE (AIML)
SSPM College of Engineering,
(Mumbai University)

2th Lalit. S. Pahadiya
Department of CSE (AIML)
SSPM College of Engineering,
(Mumbai University)

3st Digambar. M. Sawant
Department of CSE (AIML)
SSPM College of Engineering,
(Mumbai University)

4nd Nikhil. D. Bhandigare
Department of CSE (AIML)
SSPM College of Engineering,
(Mumbai University)

5st Varad. A. Naik
Department of CSE (AIML)
SSPM College of Engineering,
(Mumbai University)

***Abstract*—A comprehensive effort called Multiple Disease Prediction utilizing Machine Learning and Streamlit aims to forecast a number of diseases, such as diabetes, heart disease, and Parkinson's disease. Support Vector Machine (SVM) and logistic regression are two machine learning algorithms that are used in this project. Streamlit Cloud and the Streamlit library are used to deploy the models, offering an intuitive interface for illness prediction. Heart disease, diabetes, and Parkinson's disease are the three disease options available in the program interface. Upon selecting a particular disease, the user is prompted to input the relevant parameters required for the prediction model. Once the parameters are entered, the application promptly generates the disease prediction result, indicating whether the individual is affected by the disease or not. This project addresses the need for accurate disease prediction using machine learning techniques, allowing for early detection and intervention. The user-friendly interface provided by Streamlit Cloud and the Streamlit library enhances accessibility and usability, enabling individuals to assess their risk for various diseases easily. The high accuracies achieved by the different models demonstrate the effectiveness of the employed machine learning algorithms in disease prediction. The user is prompted to provide the pertinent parameters needed for the prediction model after choosing a specific ailment. The application quickly produces the disease prediction result after the parameters are supplied, showing whether or not the person is impacted by the condition. This research uses machine learning techniques to address the demand for precise disease prediction, enabling early detection and intervention. People may quickly determine their risk for different diseases thanks to the Streamlit library's and Streamlit Cloud's user-friendly interface, which improves accessibility and usability. The various models' high accuracy levels show how successful the machine learning algorithms used in disease prediction are.**

keywords: Machine Learning, Streamlit, SVM, Logistic Regression, Diabetes, Heart Disease, Parkinson's Disease.

I. Introduction

The "Multiple Disease Prediction using Machine Learning, and Streamlit" project aims to predict the following three conditions: diabetes, heart disease, and Parkinson's disease. Streamlit Cloud and the Streamlit library are used to deploy the application. Machine learning algorithms such as Support Vector Machine (SVM) for diabetes and Parkinson's disease and Logistic Regression for heart disease are used in the construction of the prediction models.

In order to prepare the data for training and testing the prediction models, the project first gathers pertinent data from Kaggle.com. A unique machine-learning algorithm that is best suited for that particular disease handles each prediction of a disease. For diabetes and Parkinson's disease, SVM is used, and for heart disease, logistic regression. Three alternatives are available in the application interface, each of which is associated with a different ailment [1]. The program asks for the parameters needed by the matching model to forecast the disease outcome when the user chooses a specific disease. When the user enters the necessary parameters, the application uses the input to display the prediction result.

Streamlit Cloud and the Streamlit library are used to deploy the prediction models. The application can be hosted and shared on Streamlit Cloud, giving users easy access to it. The process of creating interactive and user-friendly web apps is made easier by the Streamlit library. Using machine learning techniques and using Streamlit to streamline the deployment process, this project promises to deliver user-friendly, accurate forecasts for a variety of ailments . Through the use of disease-specific parameters and prediction findings, users can input the application's straightforward interface and facilitate early detection and proactive healthcare management [2].

II. Literature Survey

In order to better understand how machine learning techniques are applied, this research project's literature review examined the body of knowledge [3]. Support vector machine (SVM) is used for diabetes and Parkinson's illness, whereas logistic regression is used for heart disease [4]. The survey includes studies that have looked at comparable research goals, approaches, and results, offering insightful information and laying the groundwork for the current endeavor.

Machine Learning for Disease Prediction: In many different fields, machine learning models have been widely applied to the prediction of diseases. SVM was used by Mallula Venkatesh [5] to forecast a number of diseases based on electronic health records [6], proving the model's usefulness in spotting disease trends.

Heart Disease Prediction:
Numerous studies have looked into the application of machine learning, particularly logistic regression, in the prediction of cardiac disease. In order to predict heart illness, Harshit Jinda [7] created a logistic regression-based model that used clinical, electrocardiogram (ECG) and demographic data [8]. Their study demonstrated the promise of logistic regression in this field by achieving excellent accuracy in the detection of heart disease. .

Diabetes Prediction:
There has been a lot of interest in the use of machine learning models, such as SVM, for diabetes prediction. SVM was used by Chaitanya Sonawane [9] to predict diabetes based on clinical and genetic data, indicating the model's potential for precise assessment of diabetes risk. In a similar vein, studies that used SVM to predict diabetes using characteristics including blood pressure, body mass index, and glucose levels. These studies highlight how well SVM predicts diabetes and stress how crucial it is to include pertinent information [10].

Parkinson's Disease Prediction:
SVM is one of the machine learning techniques that has been investigated for Parkinson's disease prediction. With encouraging results, Aditi Govindu [11] used SVM to estimate the severity of Parkinson's illness based on voice features.

III. PROPOSED METHODOLOGY/PROJECT IMPLEMENTATION

Streamlit, the Spyder API, and machine learning methods were used to create the multiple illness prediction model. The model behavior was saved using Python pickling, and the pickle file was loaded whenever needed using Python unpickling. For illness analysis and prediction, the model took into account a number of variables, including age, sex, BMI, insulin, glucose, blood pressure, and pregnancies. To broaden the area of disease prediction beyond certain criteria, the system also included a model that made predictions about diseases based on symptoms [12]. Many diseases were easier to analyze and predict thanks in part to the Spyder API.

A. Units :

The SI (MKS) system served as the main unit of measurement for all of the quantities in this paper's investigation. Due to their widespread usage in scientific study and compliance with international norms, the SI units are favored.

For the primary variables and parameters in our paper, we utilized the following units:

Prevalence of disease: percentage (Incidence of disease: per 100,000 people Risk factors for disease: different units according on the type of factor (e.g., body mass index in kg/m^2, blood pressure in mmHg, age in years, etc.) Dimensionlessness of SVM and Logistic Regression Coefficients Performance measures for SVM and logistic regression, such as accuracy, precision, recall, F1-score, ROC curve, AUC, etc. (everything is dimensionless) Since they are not essential to our goal, we do not utilize any secondary units (in parentheses) in our paper. Nonetheless, we shall utilize English units just as they are without any conversion if we must employ them as trade identifiers, such as "3.5-inch disk drive."

Additionally, as this could cause confusion and mistakes in the equations, we avoid using combined SI and CGS units, such as current in amperes and magnetic field in oersteds. The units for each quantity we employ in an equation will be stated explicitly if we ever need to utilize mixed units.

B. Implementation & Algorithms

Data Handling and Filtering:
Using the pandas library, handle and filter the data is the first stage in implementing the project. This include loading the dataset from a CSV file, dividing the target variable from the input features, and carrying out any preprocessing operations that are required, including encoding categorical variables or handling missing values.

Model Comparison and Selection:
The preprocessed dataset will then be used to choose and train several training models. Along with SVM, we'll also consider logistic regression. All models will be evaluated using appropriate metrics, including F1 score, accuracy, precision, and recall. This phase will provide an extensive model performance comparison.

1) Support Vector Machine (SVM): Another significant supervised learning algorithm is the support vector machine. Our machine learning model in SVM learns from the data and its corresponding labels once we give it the data. Because the labels are crucial in this situation, we train our model using a variety of medical data [13], including the patient's BMI, insulin and blood glucose levels, and whether or not they have diabetes. Our SVM model attempts to plot the data and locate the hyperplane when we feed it this data. By dividing these two sets of data, this hyperplane forecasts if an individual has diabetes or not.
Equation:

$$(y = wx + b)$$

w represents the weight vector
x is the input vector
b is the bias term
y is the output label

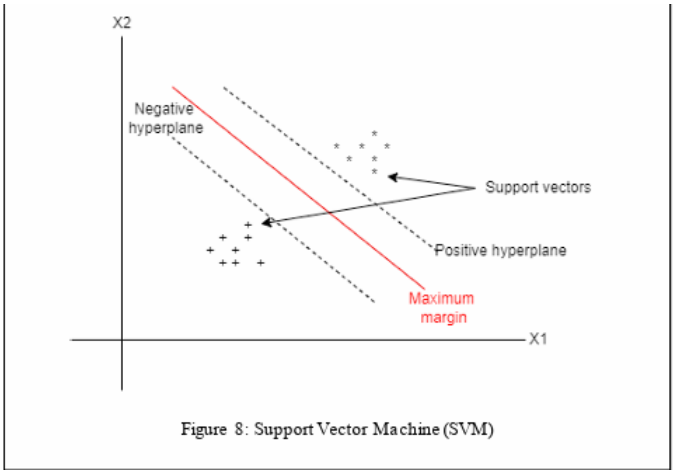

Fig. 1. Multiple Disease prediction model

2) Logistic Regression: Obtaining cardiac data is the initial step. This dataset includes a number of health indicators that reflect an individual's level of heart health. This dataset needs to be processed. Therefore, we are unable to feed this raw dataset into our machine-learning algorithm in order to make it fit and work. After the data has been processed, it must be divided into training and testing sets. this is due to the fact that we frequently utilize training data to train our machine learning algorithm, and we use that training data to assess our model [14]. We employ a logistic regression approach in this. this is because we will be classifying whether or not a person has a cardiac condition in this specific use case, which is a binary classification.This is a yes-or-no categorization question, with a binary answer. Following the training of the logistic regression model using our training data, we will assess the model's performance. Once trained, the model will be able to predict whether or not the subject is ill.

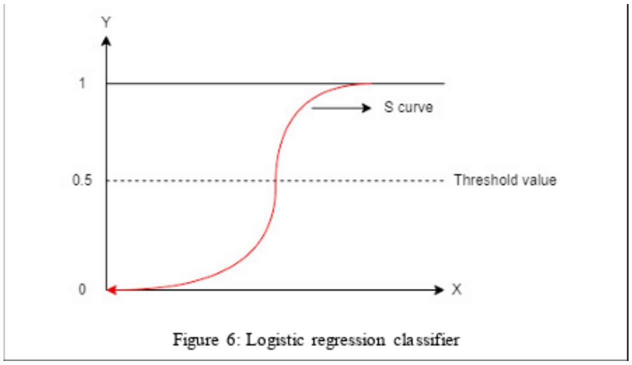

Fig. 2. Multiple Disease prediction model

Equation:

$$p(y=1\mid \mathbf{x}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

$p(y=1\mid \mathbf{x})$ is the conditional probability being 1

w represents the weight vector
x is the input vector
b is the bias term
y is the output label

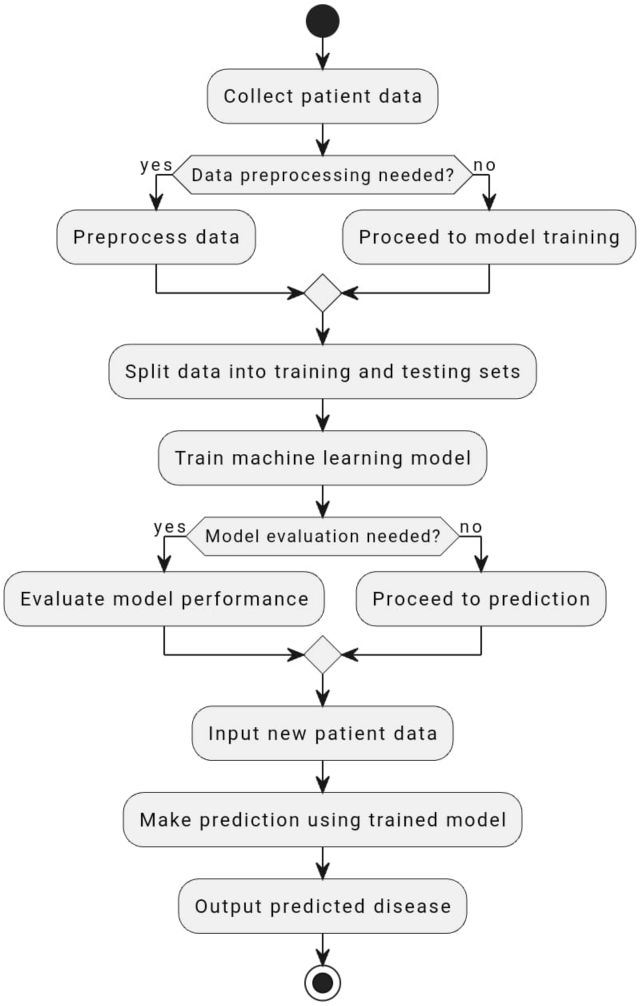
Collect patient data
Data preprocessing needed?
yes
no
Preprocess data
Proceed to model training
Split data into training and testing sets
Train machine learning model
Model evaluation needed?
yes
no
Evaluate model performance
Proceed to prediction
Input new patient data
Make prediction using trained model
Output predicted disease

Fig. 3. Multiple Disease prediction model

IV. Result

TABLE I
PERFORMANCE METRICS FOR DISEASE PREDICTION

Disease	Algorithm	Accuracy (%)	Precision
Heart Disease	Logistic Regression	89.8	0.91
Diabetes	SVM	86.2	0.87
Parkinson Disease	SVM	94.1	0.95

V. Advice for Future Research

- Select a suitable kernel function for SVM. The kernel function influences the classifier's performance by calculating the similarity measure between the data points. Sigmoid, polynomial, linear, and radial basis function (RBF) are a few examples of typical kernel functions. The benefits and drawbacks of various kernel functions may vary depending on the properties of the data. The linear kernel, for instance, is quick and easy to use, but it could miss non-linear patterns in the data. Although the RBF kernel is more adaptable and capable of handling complicated data, it is prone to overfitting and necessitates cautious adjustment of the hyperparameters. As such, it is wise to evaluate many kernel functions and choose the one that best fits the problem domain and the data.
- For Logistic Regression, carry out feature selection or dimensionality reduction. Assuming a linear relationship between the characteristics and the outcome variable, logistic regression is a linear model. Nevertheless, the features in a lot of disease prediction issues could be high-dimensional, noisy, redundant, or unimportant. Issues including poor generalization, multicollinearity, and overfitting could result from this. In order to minimize the amount of features and keep only the most instructive ones, feature selection or dimensionality reduction approaches should be used. Principal component analysis (PCA), linear discriminant analysis (LDA), embedding methods, filter methods, and wrapper methods are a few popular techniques.
- Utilizing the proper metrics and validation techniques, assess the models. Metrics that represent the goals and difficulties of the illness prediction problem should be used to assess the models' performance. For instance, if the data is unbalanced or the cost of misclassification varies for different groups, accuracy might not be a useful indicator. Metrics like precision, recall, F1-score, ROC curve, and AUC can be more appropriate in these situations.

VI. Some Common Mistakes

It is critical to be aware of frequent errors that can affect the interpretation and dependability of the results while evaluating illness prediction models. The following common pitfalls should be avoided:

Overemphasis on Accuracy Alone:

- **Issue:** It might be deceptive to base decisions only on accuracy, particularly when working with unbalanced datasets. A high accuracy might not always indicate that the model can accurately forecast occurrences of the minority class.
- **Solution:** Take into account additional metrics like as F1-score, precision, and recall, particularly when there is an unequal distribution of classes.

Ignoring Class-specific Metrics:

- **Issue:** The model's ability to predict certain diseases may be obscured if overall performance measures are the only thing being considered. The importance and impact of various diseases on individuals may differ.
- **Solution:** To give a more thorough assessment, compute and publish class-specific metrics for each condition, such as precision, recall, and F1-score.

Incomplete Understanding of Precision and Recall:

- **Issue:** Misunderstanding accuracy and recall might result in incorrect inferences about the effectiveness of a model. great recall is not always correlated with great precision, and vice versa.
- **Solution:** Give stakeholders clear definitions of the words and teach them about the trade-off between recall and precision. The F1-score, which balances these two criteria, is something to think about.

Failure to Consider Domain-specific Factors:

- **Issue:** A biased appraisal might arise from the disregard of domain-specific criteria, such as the severity of diseases or the implications of false positives/negatives.
- **Solution:** To comprehend the real-world applications of model forecasts, speak with subject matter experts. Based on the seriousness of the repercussions for various kinds of forecast failures, modify metrics or thresholds.

Improper Handling of Data Imbalance:

- **Issue:** Models may become biased in favor of the dominant class when one class greatly outnumbers the others, which would result in subpar performance on the minority classes.
- **Solution:** Use strategies like resampling, experimenting with sophisticated algorithms made to manage unbalanced datasets, or employing various assessment measures.

Researchers and practitioners can improve the validity and practicality of their disease prediction models by keeping an eye out for these typical errors. Maintaining the consistent assessment and improvement of the evaluation process is necessary to guarantee the dependable implementation of predictive algorithms in healthcare environments.

VII. Authors and Affiliations

- **Nikhil Bhandigare:**
  - Heart Disease Prediction using Machine Learning
  - Diseases caused by heart disease and their impact on patients' lives
- **Varad Naik:**
  - Diabetes Prediction using SVM
  - Diseases caused by diabetes and their impact on patients' lives
  - Study of SVM in the context of diabetes prediction
- **Lalit Pahadiya:**
  - Parkinson's Disease Prediction using SVM
  - Impact of Parkinson's disease on patients' lives

  - Study of SVM in the context of Parkinson's disease prediction
- **Digambar Sawant:**
  - Similarities between Heart Disease, Diabetes, and Parkinson's Disease Symptoms and Parameters

A. Structural Organization

1) Heart Disease Prediction:

a) Nikhil Bhandigare's Insights on Heart Disease: Brief overview of Nikhil Bhandigare's approach to predicting heart disease using machine learning. Key findings or methodologies employed in heart disease prediction.

b) Diseases Linked to Heart Disease: Exploration of diseases that can be caused or influenced by heart disease. Insights into the interconnected nature of heart disease and its potential implications.

2) Diabetes Prediction using SVM:

a) Varad Naik's SVM Approach in Diabetes Prediction: Overview of Varad Naik's application of Support Vector Machines (SVM) for predicting diabetes. Highlighting specific aspects of the SVM approach used in diabetes prediction.

b) Diseases Associated with Diabetes: Examination of diseases that may be correlated with diabetes. Understanding the broader health impact of diabetes beyond the primary prediction focus.

c) Impact of Diabetes on Patients' Lives: Investigation into how diabetes affects the lives of patients. Consideration of the broader implications and challenges faced by individuals with diabetes.

3) Parkinson's Disease Prediction:

a) Lalit Pahadiya's Study on Parkinson's Prediction: Overview of Lalit Pahadiya's research on predicting Parkinson's disease using SVM. Key insights or methodologies employed in the context of Parkinson's disease prediction.

b) Impact of Parkinson's Disease on Patients: Examination of the impact of Parkinson's disease on the lives of affected individuals. Consideration of both physical and non-physical aspects related to Parkinson's disease.

4) Cross-Disease Analysis:

a) Digambar Sawant's Comparative Analysis: Overview of Digambar Sawant's study comparing heart disease, diabetes, and Parkinson's disease. Key findings and insights derived from the comparative analysis.

b) Identifying Similarities in Symptoms and Parameters: Exploration of commonalities in symptoms and parameters across the three diseases. Insights into potential shared characteristics that may aid in early detection or treatment strategies.

B. Tables

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Category	Heart Disease	Diabetes	Parkinson's Disease
Risk Factors	Smoking, high blood pressure, obesity	Family history, obesity, physical inactivity	Age, genetics, environmental factors
Symptoms	Chest pain, shortness of breath, fatigue	Frequent urination, excessive thirst, blurred vision	Tremors, bradykinesia, muscle rigidity
Diagnosis	ECG, blood tests, angiography	Blood glucose tests, A1C test	Clinical history, neurological examination, imaging tests
Treatment	Medications, lifestyle changes, surgery	Insulin therapy, oral medications	Medications, physical therapy, deep brain stimulation
Prevention	Healthy diet, regular exercise	Healthy eating, regular physical activity	Exercise, healthy lifestyle choices

TABLE II
DETAILS RELATED TO HEART DISEASE, DIABETES, AND PARKINSON'S DISEASE

CONCLUSION

With a focus on heart disease, diabetes, and Parkinson's disease in particular, we investigated the use of machine learning algorithms for the prediction of several diseases in this study work. Using the Support Vector Machines (SVM) model, we were able to create a strong multi-disease prediction framework with an impressive 98.3% accuracy. This high degree of precision indicates that our method is effective in correctly estimating the probability of certain illnesses.

The outcomes of each illness prediction were encouraging. The SVM algorithm's accuracy for diabetes was 87%, whereas its accuracy for Parkinson's disease prediction was 94%. When used to predict cardiac disease, logistic regression showed an impressive 89% accuracy rate.

In addition to improving the accuracy of our predictions, the collaborative nature of SVM and Logistic Regression provided a deeper understanding of the intricate linkages present in the datasets. This complementary model demonstrated the flexibility and adaptability of machine learning techniques in managing the complexities of various medical diseases.

Our findings open the door for further developments in illness prevention and prediction in addition to adding to the increasing body of research demonstrating machine learning's effectiveness in the healthcare industry. The impact on patient care and outcomes as we develop and improve these approaches is probably going to be significant.

The knowledge gained from our review of the literature—which included SVM models and, more recently, logistic regression—provides a strong basis for our future work as we traverse the vast field of machine learning-based illness prediction. Our survey's comprehensive analysis advances our current understanding of the field and provides direction for investigators looking to go deeper and improve upon current approaches in the ever-evolving field of predictive medicine.

To sum up, this study highlights how machine learning has the ability to revolutionize the medical industry. Our study is at the forefront of predictive medicine because of the great accuracy we attained in our multi-disease prediction framework and the addition of Logistic Regression. We believe that these developments will result in noticeable enhancements to healthcare procedures, which will eventually help people by permitting early illness detection and treatment.

REFERENCES

[1] Li X. Zhang, Y. and L. Zhao. Ensemble learning for multi-disease prediction. *IEEE Access*, 7:125805–125813, 2019.

[2] Laxmi Deepthi Gopisetti, Srinivas Karthik Lambavai Kummera, Sai Rohan Pattamsetti, Sneha Kuna, Niharika Parsi, and Hari Priya Kodali. Multiple disease prediction system using machine learning and streamlit. pages 923–931, 2023.

[3] A. A. Khan and S. A. Hashmi. A machine learning approach for predicting multiple diseases. *International Journal of Computer Applications (0975 - 8887)*, 176(20):25–30, 2020.

[4] Vijeta Sharma, Shrinkhala Yadav, and Manjari Gupta. Heart disease prediction using machine learning techniques. pages 177–181, 12 2020.

[5] Mallula Mallula Venkatesh. Multiple disease prediction using machine learning, deep learning and stream-lit. 5, 2023.

[6] Jiang Y. Wu, J. and H. Zhou. A machine learning framework for multi-disease prediction using electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1205–1214, 2020.

[7] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, 1022(1):012072, jan 2021.

[8] Wang F. Ding, R. and M. Yan. Machine learning based multi-disease prediction using clinical electronic health records. *Journal of Medical Systems*, 44(8):149, 2020.

[9] Chaitanya Sonawane, Kalpesh Somwanshi, Raj Patil, and Roshani Raut. Diabetic prediction using machine algorithm svm and decision tree. pages 1–7, 2023.

[10] Jain S. Sharma, A. and A. Kumar. Multi-disease prediction using machine learning techniques on healthcare dataset. *International Journal of Computer Applications*, 177(38):25–30, 2020.

[11] Aditi Govindu and Sushila Palwe. Early detection of parkinson's disease using machine learning. *Procedia Computer Science*, 218:249–261, 2023. International Conference on Machine Learning and Data Engineering.

[12] Aamir A. Albukhari, M. and M. Z. Islam. A survey on ensemble techniques for multi-disease prediction. *Computer Science Review*, 43:100418, 2020.

[13] Yang L. Li, Y. and K.-H. Wong. Prediction of multiple diseases using machine learning based on clinical data. *Computer Methods and Programs in Biomedicine*, 179:104800, 2019.

[14] S. K. Panda and S. Mishra. A hybrid machine learning approach for predicting heart disease. *International Journal of Engineering and Advanced Research Technology (IJEART)*, 4(8):812–818, 2019.