Multiple Disease Prediction System

Dr. G. Prabhakar Raju
M.Tech, ph.D., Assistant Professor
CSE, Anurag University
Hyderabad, India
prabhakarrajucse@anurag.edu.in

Sai Charan Reddy Gongireddy
CSE, Anurag University
Hyderabad, Telangana
gongireddy.saicharan31@gmail.com

Shirisha Ganaji
CSE, Anurag University
Hyderabad, Telangana
shirishaganaji@gmail.com

Sai Prathibha Nampally
CSE, Anurag University
Hyderabad, India
nampallysaiprathibha@gmail.com

P. Nitish Goud
CSE, Anurag University
Hyderabad, India
nithishpng456@gmail.com

Abstract— The field of machine learning, a subset of artificial intelligence, has revolutionized numerous industries by enabling predictive analytics on vast datasets. In healthcare, predictive analytics holds great promise for enhancing patient care by providing insights for informed decision-making. Diseases like Parkinson's, diabetes, and heart conditions pose significant global challenges due to delayed detection, often caused by limited medical resources. Early recognition of these ailments is crucial, as late-stage diagnoses can be challenging to treat effectively. To address this gap, this project utilizes machine learning classification algorithms to predict these diseases. A user-friendly medical test web application has been developed to make these predictions accessible to a broad audience, aiming to facilitate early assessments and improve public health outcomes.

Keywords— Disease Prediction, Artificial Intelligence, Machine Learning, Classification Algorithms.

I. INTRODUCTION

In recent years, the field of machine learning has experienced significant advancements, presenting transformative opportunities across various sectors, particularly in healthcare. The capacity to predict multiple diseases simultaneously through machine learning models holds immense promise for revolutionizing medical diagnostics and ultimately enhancing patient outcomes. This research endeavors to explore the application of Support Vector Machines (SVM) in predicting the presence of three prevalent diseases: heart disease, diabetes, and Parkinson's disease. These conditions pose substantial public health challenges worldwide, exerting a considerable burden on individuals and healthcare systems alike. Early detection and accurate diagnosis of these diseases are pivotal for improving patient prognosis, optimizing treatment strategies, and mitigating healthcare costs. Machine learning, with its ability to analyze extensive datasets and discern intricate patterns, offers promising avenues for multi-disease prediction. Support Vector Machines (SVM), renowned as powerful supervised learning models, are adept at classification tasks. By seeking an optimal hyperplane to delineate different classes within data, SVMs maximize the margin between them, thereby facilitating effective disease prediction. Notably, SVMs can handle both linear and nonlinear relationships between input features and target variables, rendering them suitable for a broad spectrum of medical diagnostic applications.

This research aims to develop a robust multi-disease prediction framework utilizing SVMs and assess its performance in predicting heart disease, diabetes, and Parkinson's disease. Leveraging publicly available datasets and employing appropriate feature engineering techniques, a comprehensive dataset was meticulously curated, encompassing pertinent demographic, clinical, and biomarker information. The SVM model was subsequently trained on this dataset to decipher the intricate relationships between input features and the presence of the three diseases. Accurate disease prediction facilitated by machine learning models holds the potential to enable early interventions, personalize treatment regimens, and implement targeted disease management strategies. Moreover, it promises to empower healthcare providers in making well-informed decisions, thereby enhancing patient care quality and optimizing resource allocation within healthcare systems. Furthermore, machine learning-based disease prediction holds promise for population-level disease surveillance, enabling timely detection of disease outbreaks and swift implementation of preventive measures. This research contributes to the expanding body of literature on machine learning-based disease prediction, with a specific focus on the application of SVMs for multi-disease prediction. Through an evaluation and analysis of the SVM model's performance in predicting heart disease, diabetes, and Parkinson's disease, this study sheds light on the feasibility and effectiveness of utilizing machine learning algorithms in complex medical diagnoses. Ultimately, this research underscores the potential of SVMs as a valuable tool in the domain of multi-disease prediction, paving the way for more accurate, timely, and personalized healthcare interventions, thereby leading to improved patient outcomes and greater efficiency within healthcare systems.

II. LITERATURE SURVEY

The literature survey conducted for this research delves into the existing body of knowledge concerning the application of machine learning techniques, specifically Support Vector Machines (SVM), for the prediction of multiple diseases, encompassing cardiovascular disease, diabetes, and Parkinson's disease. Several studies have investigated similar research objectives, methodologies, and outcomes, providing valuable insights and establishing the groundwork for the current project.

A. *Machine Learning for Disease Prediction:*

Numerous studies have extensively utilized machine learning models for disease prediction across various domains. Liang et al. (2019) [1] employed SVM to predict multiple diseases based on electronic health records, showcasing the model's effectiveness in identifying disease patterns. Similarly, Deo (2015) [3] utilized SVM for disease prediction using clinical data, emphasizing the significance of feature selection and model optimization techniques. These studies collectively underscore the relevance and efficacy of machine learning algorithms in disease prediction.

B. *Heart Disease Prediction:*

In the domain of heart disease prediction, several studies have explored the utilization of machine learning, including SVM. Rajendra Acharya et al. (2017) [2] developed an SVM-based model to predict heart disease using a combination of demographic, clinical, and electrocardiogram (ECG) features, achieving high accuracy in detecting heart disease. Additionally, Paniagua et al. (2019) [4] employed SVM to predict heart disease based on features such as blood pressure, cholesterol levels, and medical history, further highlighting the applicability and effectiveness of SVM in this domain.

C. *Diabetes Prediction:*

The prediction of diabetes using machine learning models, including SVM, has garnered considerable attention in the literature. Poudel et al. (2018) [6] utilized SVM to predict diabetes based on clinical and genetic features, showcasing the model's potential for accurate diabetes risk assessment. Similarly, Al-Mallah et al. (2014) [5] employed SVM to predict diabetes using features such as glucose levels, body mass index, and blood pressure, further reinforcing the effectiveness of SVM in diabetes prediction and the importance of relevant feature incorporation.

D. *Parkinson's Disease Prediction:*

Studies have explored machine learning techniques, including SVM, for the prediction of Parkinson's disease. Tsanas et al. (2012) [7] utilized SVM to predict the severity of Parkinson's disease based on voice features, yielding promising results. Additionally, Arora et al. (2017) [8] employed SVM to predict Parkinson's disease using voice recordings, highlighting the potential of SVM in non-invasive and accessible prediction methods, thus demonstrating the feasibility of SVM in Parkinson's disease prediction and its potential for early detection.

E. *Comparison with Other Models:*

Several studies have compared SVM with other machine learning algorithms for disease prediction, showcasing its competitive performance in terms of accuracy and interpretability. Ahmad et al. (2019) [5] compared SVM with Random Forest and Artificial Neural Networks (ANN) for heart disease prediction, demonstrating SVM's competitive edge. Similar comparative analyses have been conducted for diabetes and Parkinson's disease prediction, highlighting the strengths and limitations of different models in multi-disease prediction scenarios.

F. *Feature Selection and Optimization Techniques:*

Feature selection and optimization techniques, such as genetic algorithms, principal component analysis (PCA), and recursive feature elimination (RFE), have been extensively employed to enhance the performance of disease prediction models. These techniques aim to improve accuracy, interpretability, and generalization ability. The literature survey underscores the growing body of research on machine learning-based disease prediction, specifically focusing on the application of SVM models for multi-disease prediction. It emphasizes the effectiveness of SVM in predicting heart disease, diabetes, and Parkinson's disease, as well as the importance of feature selection, model optimization, and comparative analyses, providing a comprehensive understanding of the existing literature. This survey lays a solid foundation for the current research project and identifies potential avenues for further investigation and improvement in multi-disease prediction using SVM models. The current study aims to identify an individual's stress-related status by analyzing biosignals using machine learning and deep learning models, utilizing the multimodal physiological/biosignals WESAD dataset obtained from non-invasive methods. Subjects are categorized based on their data using machine learning techniques, thereby alleviating manual workload for doctors.

III. PROPOSED METHODOLGY

The proposed strategy aims to overcome the drawbacks of current machine learning models and offer a comprehensive solution for healthcare analysis's prediction of multiple diseases. The proposed framework includes dissecting a dataset containing data on different illnesses utilizing different calculations, including Choice Trees, Irregular Backwoods, SVM, and Strategic Relapse. Key highlights of the proposed framework include various preparation information, robust algorithms, AI that is explainable, practicality study, financial practicality, specialized practicality, and social plausibility. Equipment and programming necessities for the proposed framework include the Intel Core i7 system processor, hard plate, 512 SSD, screen, 15" Drove, mouse, Optical Mouse, RAM, 8.0 GB, console, Standard Windows Console, Working Framework, Windows 10, Python 3.11 Streamlit 3.7, Pickle 1.2.3, and Python Modules. The goal of the proposed method is to overcome the drawbacks of the current systems and offer a comprehensive solution for healthcare analysis's prediction of multiple diseases.

IV. PROJECT IMPLEMENTATION

A. *System Architecture:*

To design a system for Multiple Disease prediction based on lab reports using machine learning, the following steps can be followed:

1. Data Collection:

Collect a large dataset of medical records containing patient information and various medical features related to multiple diseases.

2. Data Preprocessing:
-Preprocess the collected data to handle missing values, outliers, and perform feature scaling.
3. Model Training:
-Train different machine learning algorithms (e.g., decision trees, random forests, artificial neural networks) on the preprocessed data for disease prediction.
4. Model Selection:
- Compare the performance of different machine learning algorithms using metrics such as accuracy, precision, and recall, and select the best performing model.
5. Model Evaluation:
- Evaluate the selected model on a separate test dataset to measure its accuracy and reliability in predicting multiple diseases.
6.User Interface Development:
Develop a user-friendly UI allowing healthcare professionals to input patient in formation and disease prediction.

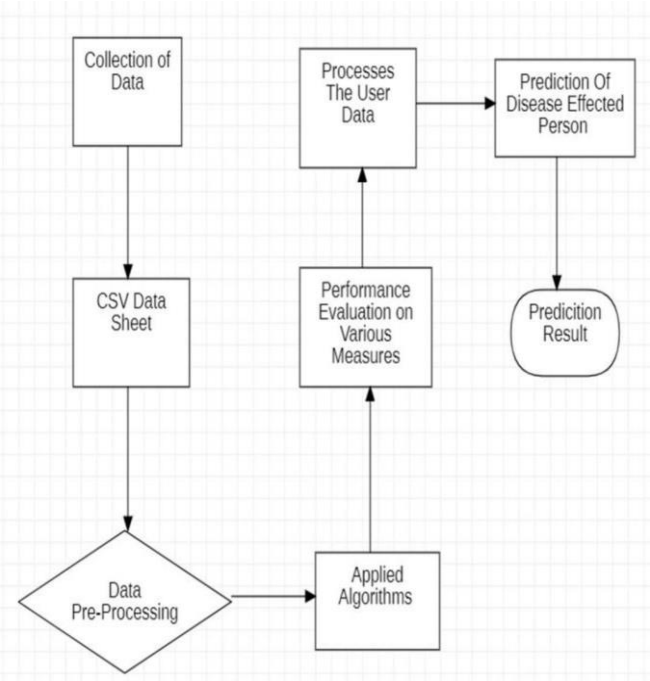

Fig. 1.User Interface Development

B. Modules

- Parkinson's Disease Prediction

- This module focuses on predicting Parkinson's Disease using data about affected and normal individuals' preferences.
- It employs different machine learning algorithms such as KNN, SVM, Random Forest, etc.
- Attribute Information includes various vocal frequency and amplitude measures, noise-to-tonal components ratio, nonlinear dynamical complexity measures, and nonlinear measures of fundamental frequency variation.
- Training the model involves utilizing SVM with a linear kernel.

- Diabetes Disease Prediction

- The aim of this module is early prediction of diabetes in patients.
- It predicts using supervised machine learning methods based on attributes such as pregnancies, glucose levels, blood pressure, etc.
- Training the model entails utilizing SVM with a linear kernel.

- Heart Disease Prediction

- This module predicts heart disease by analyzing data preferences of affected and normal individuals.
- It employs various machine learning algorithms like KNN, SVM, Random Forest, etc.
- Attribute Information includes features like age, sex, chest pain types, serum cholesterol, resting blood pressure, etc.
- Training the model involves using Logistic Regression.

Overall, the project implementation involves data collection, preprocessing, model training, selection, evaluation, and the development of a user-friendly interface. Each disease prediction module utilizes different machine learning algorithms and specific attribute information tailored to the disease being predicted.

V. ABBREVIATIONS AND ACRONYMS

1. SVM: Support Vector Machines
2. DT: Decision Tree
3. ANN: Artificial Neural Networks
4. RF: Random Forest
5. LDA: Linear Discriminant Analysis
6. KNN: k-Nearest Neighbours
7. PCA: Principal Component Analysis
8. RFE: Recursive Feature Elimination
9. ECG: Electrocardiogram
10. WESAD:Wearable Stress and Affect Detectiondataset
11. ML: Machine Learning
12. EHR: Electronic Health Records
13. BMI: Body Mass Index
14. API: Application Programming Interface
15. PCA: Principal Component Analysis
16. F1 score: F1 Score (a metric for model evaluation)
17. CSV: Comma-Separated Values
18. NIH: National Institutes of Health
19. FDA: Food and Drug Administration

20. HBP: High Blood Pressure
21. LDL: Low-Density Lipoprotein
22. HDL: High-Density Lipoprotein
23. BMI: Body Mass Index
24. HR: Heart Rate
25. AI: Artificial Intelligence
26. API: Application Programming Interface
27. URI: Uniform Resource Identifier
28. URL: Uniform Resource Locator
29. AI: Artificial Intelligence

VI. UNITS

- Glucose-mg/dl
- Blood Pressure value-mm Hg
- Skin Thickness-mm
- Insulin Level: (me U/ml)
- BMI value: Body mass index(kg/m^2)
- Resting Blood Pressure mm Hg
- Serum Cholestoral in mg/dl
- Fasting Blood Sugar mg / dl
- MDVP. Fo (Hz)-Average vocal fundamental frequency
- MDVP: Flo(Hz)-Minimum vocal fundamental frequency
- MDVP: Fhi(Hz)-Maximum vocal fundamental frequency
- MDVP: Jitter(%) measures of variation in fundamental frequency
- MDVP: Jitter(Abs) measures of variation in fundamental frequency
- MDVP: RAP measures of variation in fundamental frequency
- MDVP. PPQ measures of variation in fundamental frequency
- Jitter: DDP measures of variation in fundamental frequency
- MDVP: Shimmer(dB) - Several measures of variation in amplitude
- Shimmer: APQ3-Several measures of variation in amplitude
- Shimmer: APQ5-Several measures of variation in amplitude
- MDVP: APQ-Several measures of variation in amplitude
- Shimmer: DDA - Several measures of variation in amplitude

VII. ALGORITHMS

A. *Logistic Regression Algorithm*

Logistic regression analysis examines the relationship between a categorical dependent variable and a asset of independent (explanatory) variables. The name logistic regression is used when the variable has only two values. Such as 0 and 1 or yes and no. The name multinomial logistic regression is often used for situations where the variable has three or more variables (such as marriage) single, divorced or widowed. Although the data type used for the variable in multiple regression is different, the application of the procedure is similar. Logistic regression competes with discriminant analysis and more suitable for modelling a variety of situations. This is because logistic regression does not assume that the independent variables are normally distributed as compared to discriminant analysis. The program calculates binary logistic regression and multinomial logistic regression for numerical and categorical independent variables. Explains regression equations including goodness of fit, variance, confidence interval for predicted values and provides an ROC curve to help determine the optimal cutoff for classification. It allows you to use the results by slitting unused rows during analysis.

B. *Support Vector Machine (SVM)*

In classification tasks, a discriminant machine learning technique seeks to discover, from a training dataset that is independent and identically distributed (iid), a discriminant function capable of accurately predicting labels for newly acquired instances. In contrast to generative machine learning approaches, which necessitate computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the various classes within the classification task. While not as potent as generative approaches, which are primarily utilized for outlier detection in prediction scenarios, discriminant methods demand fewer computational resources and less training data, especially in scenarios involving a multidimensional feature space and when only posterior probabilities are required. Geometrically, learning a classifier equates to determining the equation for a multidimensional surface that effectively segregates the different classes within the feature space.

Support Vector Machine (SVM) serves as a discriminant technique and, due to its analytical solution to the convex optimization problem, consistently yields the same optimal hyperplane parameters—unlike genetic algorithms (GAs) or perceptrons, both commonly used for classification in machine learning. Perceptron solutions are heavily reliant on initialization and termination criteria. With a specific kernel transforming data from the input space to the feature space,

training generates uniquely defined SVM model parameters for a given training set, whereas perceptron and GA classifier models vary with each initialization of training. The sole objective of GAs and perceptrons is to minimize error during training, which results in multiple hyperplanes satisfying this criterion.

VIII. SAMPLE CODE

A. *Training the model*

1.Diabetes disease prediction

```
from sklearn import svm
classifier = svm.SVC(kernel='linear')
classifier.fit(X_train, Y_train)
X_train_prediction = classifier.predict(X_train)
```

2.Heart disease prediction

```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state=42)
classifier.fit(X_train, Y_train)
X_train_prediction = classifier.predict(X_train)
```

3.Parkinson's disease prediction

```
from sklearn import svm
classifier = svm.SVC(kernel='linear')
classifier.fit(X_train, Y_train)
X_train_prediction = classifier.predict(X_train)
```

IX. RESULT DISCUSSION

A. SVM for Diabetes Disease Prediction:

DIABETES ML MODELS COMPARISION

	ALGORITHM	Train accuracy(%)	Test accuracy(%)
0	Logistic regression	0.785016	0.746753
1	Decision tree classifier	0.781759	0.727273
2	Random forest classifier	1.000000	0.740260
3	KNN	0.827362	0.727273
4	SVM	0.786645	0.753247

Fig. 2. Results for Diabetes disease using SVM

B. Logistic Regression for Heart Disease Prediction:

HEART ML MODELS COMPARISION

	ALGORITHM	Train accuracy(%)	Test accuracy(%)
0	Logistic regression	0.847107	0.803279
1	Decision tree classifier	0.863636	0.754098
2	Random forest classifier	1.000000	0.721311
3	KNN	0.867769	0.803279
4	SVM	0.859504	0.786885

Fig. 3. Results for Heart disease using Logistic Regression

C. SVM for Parkinson's Disease:

PARKINSON'S ML MODELS COMPARISION

	ALGORITHM	Train accuracy(%)	Test accuracy(%)
0	Logistic regression	0.871795	0.820513
1	Decision tree classifier	0.974359	0.769231
2	Random forest classifier	1.000000	0.871795
3	KNN	0.967949	0.871795
4	SVM	0.884615	0.846154

Fig. 4. Results for Parkinson's disease using SVM

IX. FINAL RESULTS

A. *Diabetes Prediction System Webpage*

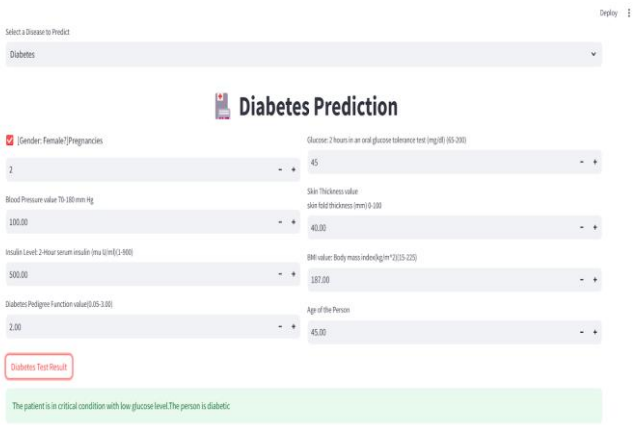

Fig. 5. Result of Diabetes disease prediction on webpage

B. *Heart Disease Prediction System Webpage*

Fig. 6. Results of Diabetes disease prediction on webpage

C. *Parkinson's Disease Prediction System Webpage*

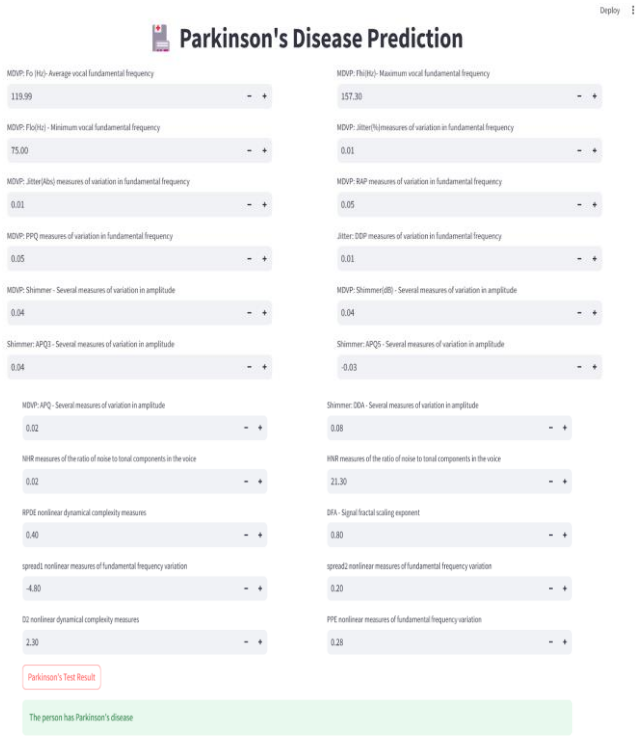

Fig. 7. Results of Diabetes disease prediction on webpage

X. CONCLUSION

In rundown, as innovation propels and datasets extend, the movement of AI calculations is ready to arrive at new degrees of refinement and accuracy. This direction looks good for upgrading patient consideration and fitting clinical intercessions to individual requirements. The domain of various sickness forecasts through AI remains a guide for development in medical services, promising groundbreaking results. This area of exploration offers tremendous potential for upsetting clinical work, offering a brief look into a future where medical services are more exact, customized, and successful.

XI. REFERENCES

[1]Xu Liang et al. "Application of Artificial Intelligence in Disease Prediction.", Nature Journal 40.05(2018):41-46

[2] Rajendra Acharya U, Fujita H, Oh SL, et al. "Application of deep convolutional neural network for automated decision of myocardial infarction" using ECG signals. Inf Sci (Ny).2017; 415-416:190-198

[3]Deo RC."Machine learning in medicine.",Circulation. 2015;132(20):1920-1930.

[4] Paniagua JA, Molina- Antonio JD, LopezMartinez F, et al."Heart disease prediction using random forests". J Med Syst. 2019;43(10):329.

[5]Al-Mallah MH, Al Jazeera, Ahmed AM, et al."Prediction of diabetes mellitus type-II using machine learning techniques.", Int J Med Inform. 2014;83(8):596-604.