Multiple Disease Prediction System

***Abstract*— Effective treatment of disease including diabetes, heart disease, and Parkinson's depends on early detection. Traditional methods can take a long time and result in a delayed diagnosis. The goal of this research is to create a machine learning-based system that can accurately predict diseases. A Support Vector Machine (SVM) was initially used, but its accuracy was poor. So Convolutional Neural Network (CNN) was used to improve performance, which achieved significantly higher accuracy in both training and testing. The results show CNN's effectiveness in medical diagnostics, providing a dependable method for early detection. Future research can focus on expanding datasets and real-time clinical applications.**

Keywords—Disease detection, Machine Learning, Support Vector Machine(SVM) Deep Learning, Convolutional Neural Network (CNN), Logistic Regression(LR), Disease Prediction.

I. INTRODUCTION

Early Detecting diseases like diabetes, heart diseases, and Parkinson's helps with better treatment and recovery. Today, doctors use tests and their knowledge. But these approaches can be expensive, time-consuming, and sometimes inaccurate. Computers can now analyze huge amounts of data and improve diagnosis thanks to artificial intelligence and machine learning. By providing quicker and more effective techniques for predicting diseases based on patient data, machine learning algorithms show significant potential in the healthcare industry. However, selecting their right model is the right way to achieve high accuracy. Support Vector Machines (SVM) were first used in this work however their accuracy was poor. To improve accuracy, Convolutional Neural Networks (CNN), a deep learning technique, were used, and the result was noticeably improved performance. The goal of this project is to create an accurate disease detection system that will help doctors make decisions and diagnose patients early. This system improves patient care by utilizing deep learning techniques to give an automated and effective method of medical diagnostics, hence decreasing reliance on manual interpretation.

II. LITERATURE REVIEW

A. *Daibetes Literature*

Iyer et al. [1] has carried out a research to predict disease diabetes applying decision tree and Naive Bayes. Diseases develop when the body does not produce enough insulin, or when the body cannot properly utilize insulin. In this work Pima Indian diabetes data set has been used. Several experiments were conducted by means of the WEKA data mining tool. In this data-set 70-30 split predict better than cross validation. For J48 74.8698% accuracy is there by using Cross Validation 76.9565% accuracy by using Percentage Split. Naive Bayes achieves 79.5652% accuracy with PS. Percentage split test gives highest accuracy when we use algorithms.

Meta-learning algorithms for diabetes disease diagnosis was presented in Sen and Dash [2]. Work data set is Pima Indians diabetes which can be obtained from UCI Machine Learning repository. WEKA is used for analysis. The patient has diabetes or not the prediction is given by the CART Adaboost Logiboost grading learning algorithms. Experiments are compared based on their correctly as well as wrongly classification. CART offers 78.646% accuracy. The Adaboost reaches 77.864% precision. Logiboost score is 77.479%. The accuracy of grading is 66.406%. B. CART CART gives the highest accuracy of 78.646% and misclassification Rate 21.354%, which is less from other techniques.

An experimental work to predict diabetes disease is done by the Kumari and Chitra [3] . Machine learning technique that is used by the scientist in this experiment is SVM. RBF kernel is used in SVM for the purpose of classification. Pima Indian diabetes data set is provided by machine learning laboratory at University of California, Irvine. MATLAB 2010a are used to conduct experiment. SVM offers 78% accuracy.

The value of the experimental on predicting of diabetes disease is performed by the Kumari and Chitra [3]. The machine learning method employed by the scientist in this experiments is the SVM. Here RBF kernel of SVM is applied for classification purpose. The Pima Indian diabetes data set Data Source : machine learning lab at the university of California, Irvine. Experiment are done with MATLAB 2010a. SVM offers 78% accuracy.

Aiswaryaet al. [5] proposes to get the solutions of detection of diabetes by exploring and analyzing the patterns which are generating from data through classification analysis using Decision Tree and Naïve Bayes techniques. The research aims to develop a simple test to detect and identify the disease in yet-to-develop stage, at a quicker pace, that would lead to early cure of the patients. Results A study has been conducted on PIMA dataset by means of cross validation approach has been carried out, gave conclusion that J48 algorithm accuracy rate is 74.8% and on the same dataset, naïve Bayes accuracy is 79.5% by using 70:30 split.

B. *Heart Literature*

Otoom et al. [6] introduced an approach for analysis and monitoring. The proposed system is used for the detection and monitoring of coronary artery disease. Cleveland heart is from UCI. The set contains 303 instances and 76 attributes. 76 features, 13 active. Detection: Two tests are applied 1) Bayes Net, 2) Support vector machine and 3) Functional Trees FT. Detection is performed using WEKA tool. Results For holdout test, we achieved 88.3% accuracy using SVM method. In Cross Validation test the Accuracy

of SVM and Bayes net are 83.8%. 81.5\% accuracy is achieved after applying FT. 7 columns are marked in nature by using Best First selection algorithm. To conduct the test on selected best 7 feature, bayes net perform 85.% of correctness, SVM achieve 86% accuracy and FT define 84% correctly.

Vembandasamy et al. [7] proposed the diagnostic method of heart disease by using Naive Bayes algorithm. Naive Bayes makes use of Bayes theorem. Hence,NBs have a strong independence assumption. The data-set used are collected from the one of the renowned diabetic research institute forms at Chennai. There are 500 patients in data set. In the Weka, the 70% of Percentage Split is applied for classification. Conversely, the accuracy of Naive Bayes is 86.419%.

Data mining techniques have been recommended by Chaurasia and Pal [8] for heart disease diagnosis. The WEKA data mining tool is utilized which consists of a collection machine learning algorithms for mining. For this perspective, Naive Bayes, J48 and bagging are employed. UCI machine learning laboratory gives the data of heart disease which have 76 attributes. The prediction is based on 11 features only. Naive bayes is 82.31% accurate. For example, J48 has the 84.35% of accuracy. The proof of proposed decision mechanism's accuracy is given in the next subsection. 85.03% Classification accuracy is provided by Bagging. Bagging has a superior classification rate on this dataset.

Parthiban and Srivatsa [9] lackluster for the diagnosis of heart disease in diabetic patient using machine learning technique. WeKA is used to apply Naive Bayes and SVM algorithms. Data set of 500 patients are considered which are gathered from Research Institute of Chennai. There are 142 patients with the disease and 358 patients who do not have the disease. With the Naive Bayes Algorithm 74% of accuracy is achieved. The optimal accuracy of 94.60 is offered by the SVM.

Tan et al. [10] introduced hybrid method where two machine-learning tools called Genetic Algorithm (G.A) and Support Vector Machine (SVM) are optimally linked through the wrapper. In the present study, we use LIBSVM and WEKA data mining tool. Model of the Experiment Five datasets (Iris, Diabetes disease, disease of Breast Cancer, Heart and Hepatitis disease) are selected from UC Irvine machine learning repository for this experiment. 84.07% for heart disease is achieved after using GA and SVM hybrid method. For diabetes data set 78.26 % accuracy is obtained. It has an accuracy of 76.20% for Breast cancer. For hepatitis disease the obtained accuracy is 86.12%.

C. Parkinson's Literature

Abiyev and Abizade (2016) [11] developed a Fuzzy Neural System (FNS) that integrates fuzzy logic and neural networks for diagnosing Parkinson's Disease. Takagi-Sugeno-Kang (TSK) fuzzy rules were constructed using Gaussian membership functions and trained with speech data from the UCI Machine Learning Repository. They applied algorithm for model training such as Decision tree, SVM, Regression and FNS. They achieved the accuracy in Decition tree(84.3%), regression(88.6%), SVM (93.84%) and FNS (99%).

In the study by Srieam et al. (2015) [12], it was shown that PD dataset has parallel dimensions more. The best prediction accuracy of 88.9 % was by SVM, followed by majority voting and KNN. Under both schemes, NB had the lowest success chunk matching rate at 69.23%, while that of RF was 90.26%. Hierarchical clus-tering and self-organizing maps (SOM) were employed for prediction, and finding more clusters for healthy datasets and fewer for diseased datasets.

Shamrat et al. (2019) [13] designed AI tools for identification of PD based on different data sources. They used SVM, KNN and LR to predict PD. Classifiers were assessed in terms of recall, precision, F1-score and accuracy. The SVM was remarkably superior and reached an accuracy of 100 % in PD prediction, while the LR generated an accuracy of 97 %. In contrast, for PD datasets KNN presented a much lower precision rate of 60 %. Findings The study demonstrated SVM as a powerful classifying algorithm in the analysis of PD datasets and indicated ML as a new approach in clinical research.

Lahmiri et al. (2019) [14] To identify voice problem patterns for diagnosing PD. They developed 8 pattern algorithms and nonlinear svm Classifier for identify individuals with pd and healthy individuals. 93% accuracy was achieved by their models.

Senturk et al. (2020) [15] introduced an automatic method for PD diagnosis based on a machine learning paradigm with feature selection and categorization. Feature selec-tion methods adopted feature importance and recursive feature elimination. The investigation considered classification trees, NNs and SVMs, among which SVMs with an RFE are superior to the other ones. The accuracy of the obtained PD diagnosis was 93.84 %, concentrating on vocal and clinical characteristics.

Gunduz et al. (2019) [16] using CNN and analyze uci speech data it provides a PD classification approach. To Reach overall model accuracy at 87% they combine feature and models.

III. METHODOLOGY

A. Diabetes Disease

In diabetes detection system dataset contains many features like pregnancies as number of pregnancies, Glucose as plasma glucose concentration, Blood pressure is diastolic blood pressure in mm Hg, Skin thickness triceps skinfold thickness in mm *Insulin as* 2-hour serum insulin in mu U/ml, *BMI* as body mass index, *Diabetes Pedigree Function as* a score estimating diabetes likelihood and *Age* as patient's age. The outcome define what are the percent as diabetes or not 0 = No, 1 = Yes. Mostly future is in integers except bmi

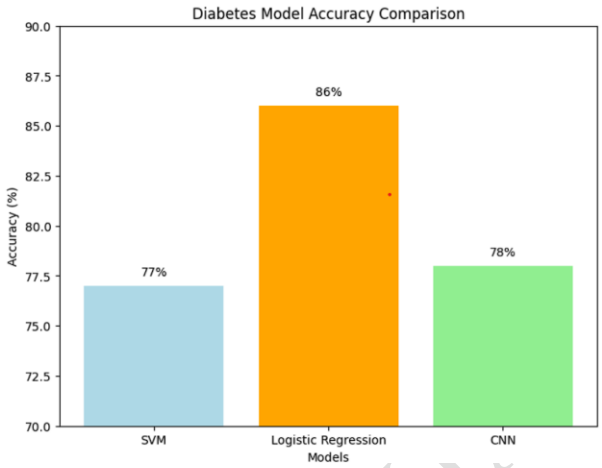
and diabetes pedigree function which are float which are float An outcome which is binary.

In Diabetes prediction system there are three machine learning algorithms use like Logistic regression, support vector machine, Convolutional Neural Network (CNN) . Each algorithm has different methods for Pre processing data, building models and making predictions based on the datasets.

Convolutional Neural Network (CNN) besed deep learning model was first use. Data loading was performed with pandas library and exploratory data analysis (EDA) was also done. StandardScaler Was used for data normalization. The feature were reshape to a 3D format Conv1D layers. In model structure Conv1D layers with 64 and 32 filters followed by BatchNormalization and Dropout layers to prevent overfitting. For binary classification flatten layer and dense layer was done using sigmoid activation function. Adam Optimizer was used for model training. Evaluation matrix use for training and testing accuracy. After testing the model we got 78% accuracy. Based on input data model can predict patient has diabetes or not.

In the second step logistic regression algorithm was used for model training with focusing on future engineering and data preprocessing. For improved data quality class-wise median imputation used for replaced missing values(zeros). Various new features were engineered like BMI values divided into six classes (Underweight to Obesity Class III), insulin were labeled as "Normal" or "Abnormal" and glucose values were binned into four bins. One-hot encoding applied to transform categorical data into numerical format. RobustScaler and StandardScaler used for preprocessing. Data was split into train-test for model training and Logistic regression model was train and tested on accuracy score and confusion matrix. After testing the model we got 86% accuracy. Based on input data model can predict patient has diabetes or not.

In the third stage an SVM-based classification model used. It start with data split into features (X) and target (Y). The features were scaled using MinMaxScaler. A Support Vector Machine (SVM) was trained using GridSearchCV to find the best hyperparameters over a range of kernels, regularization parameters, and gamma values. Using 10-fold cross-validation best performing model was chosen. To compare performance for benchmarking purposes Gradient Boosting and Random Forest classifiers were also trained and compared. After the testing model we got SVM test accuracy 69%, gradient boosting test accuracy 75% and random forest test accuracy 74%. Based on input data model can predict patient has diabetes or not.



Diabetes Accuracy Graph

B. *Heart Disease*

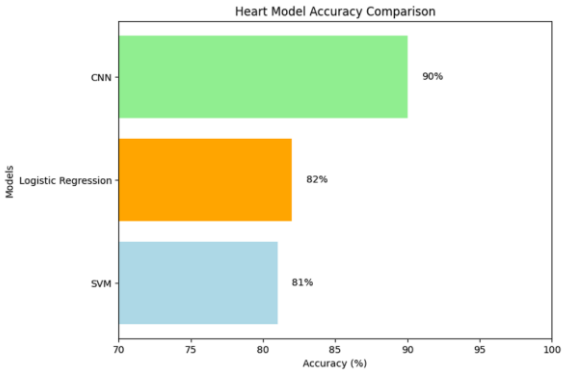
In heart disease prediction dataset include *age*, *sex* as gender, *cp* as chest pain type, *trestbps* as resting blood pressure, *chol* as serum cholesterol, *fbs* as fasting blood sugar >120 mg/dl, *restecg* as resting ECG results, *thalach* as maximum heart rate achieved, *exang* as exercise-induced angina *oldpeak* as ST depression induced by exercise, *slope* as slope of peak ST segment, *ca* as number of major vessels colored by fluoroscopy and *thal* as thalassemia type. Target variable Identify that patient has heart disease or not 0 = No, 1 = Yes. There are all type of data available in data set like integer float categorial and binary values.

In Heart prediction system there are three machine learning algorithms use like Logistic regression, support vector machine, Convolutional Neural Network (CNN) . Each algorithm has different methods for Pre processing data, building models and making predictions based on the dataset.

The first, Logistic regression algorithm was used for model training . Data loading was performed with pandas library and exploratory data analysis (EDA) was also done to understand the structure and contents of the data. The dataset was then separated into input features (X) and the target variable (Y) which indicated the presence of heart disease. Maintaining the class balance using stratification 80/20 ratio applied for train-test split data. The Logistic Regression model from scikit-learn was trained using default parameters and achieved an accuracy of 82% on the test set. Based on input data model can predict patient has heart disease or not.

In the second step Support Vector Machine (SVM) with a linear kernel was used. Data loading was performed with pandas library and exploratory data analysis (EDA) was also done to understand the structure and contents of the data. The dataset was then separated into input features (X) and the target variable (Y). This SVM model was kept simple and did not require any advanced data preprocessing. The SVM model perform well and achieve 81% accuracy on test set. The trained model give prediction that patient has heart disease or not based on input data.

In the third stage Convolutional Neural Network (CNN) algorithm used for model training. Deep neural network with dense layers instead of convolutional. Before training advanced preprocessing was applied. For noise reduction outliers in numerical data were removed using the Interquartile Range (IQR) method. StandardScaler was used to normalize numerical features and OneHotEncoder was used to encode categorical variables in order to ensure consistent scale across inputs.. Model architecture included dense layer with 64 neurons and ReLU activation followed by a dropout layer second dense layer with 32 neurons and another dropout and a final output layer with a single neuron using sigmoid activation for binary classification. Model performance was print using metrics such as accuracy, precision, recall, and F1-score. This model achieve high accuracy 90% in compare to svm and Logistic regression. As input data it take both categorical and numerical input and give prediction based on the input data that patient has heart disease or not.



Heart Accuracy Graph

C. *Parkinson's Disease*

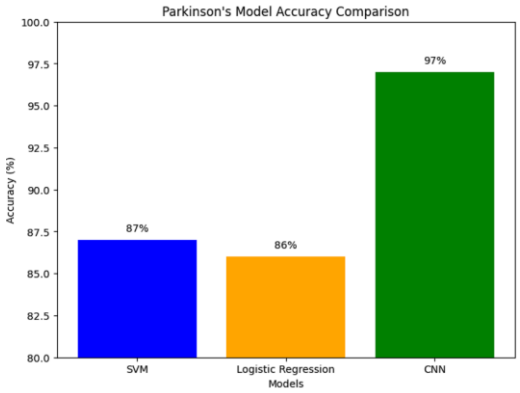
In partition precious prediction data set include future such as *MDVP:Fo(Hz)* as average fundamental frequency, *MDVP:Fhi(Hz)* as maximum vocal frequency, *MDVP:Flo(Hz)* as minimum vocal frequency, *MDVP:Jitter(%) and MDVP:Shimmer* as measures of pitch variation and amplitude variation, *HNR* as harmonics-to-noise ratio, *RPDE* as nonlinear dynamical complexity, *DFA* as signal fractal scaling exponent and *PPE* as pitch period entropy. The target variable *status* indicate the presence of Parkinson's disease in patient or not 0 = healthy, 1 = Parkinson's. Most feature are flot and target is binary.

In parkinson's prediction system there are three machine learning algorithms use like Logistic regression, support vector machine, Convolutional Neural Network (CNN) . Each algorithm has different methods for Pre processing data, building models and making predictions based same on the dataset.

The first, Logistic regression algorithm was used for model training . Data loading using pandas library and exploratory data analysis (EDA) was also done to understand the dataset structure. while class distribution was checked to ensure balance between classes. Dataset was split into training and testing sets with 80/20 ratio split. The Logistic Regression model was trained using scikit-learn. After training both training and test data were used to check the model accuracy. After testing the model we got 86% accuracy. Based on input data model can predict patient has Parkinson's disease or not.

The second model was used as support vector machine with linear kernel. It use same dataset as used by logistic regression. Data loading using pandas library and exploratory data analysis (EDA) was also done to understand the dataset structure. The svm model was trained using 80/20 train-test split of dataset to maintain class balance. After training both training and test data were used to check the model accuracy. The svm model perform 87% accuracy on test data. Based on input data model can predict patient has Parkinson's disease or not.

The third most advanced model Convolutional Neural Network (CNN) used on same dataset. It use 1D Convolutional Neural Network to identify Parkinson's disease based on data. .Data loading using pandas library and exploratory data analysis (EDA) was use for understanding data structure. In preprocessing task StandardScaler use for normalize the feature values The feature matrix reshaped to match expected 3D input shape because CNNs require structured input (samples, features, 1). The CNN consisted of Conv1D layers, batch normalization, dropout layers and dense layers ending in a sigmoid activation for binary classification and also Adam optimizer was used for improve optimization. Batch Normalization and a Dropout layer to reduce overfitting. The CNN model performance was print using metrics such as accuracy, precision, recall, and F1-score. It achieve 97% accuracy which are higher than svm and logistic regression models. Based on input data model can predict patient has Parkinson's disease or not.



Parkinson's Accuracy Graph 1

D. *Comparison Table*

Feature	Diabetes Detection	Heart Detection	Parkinson's Detection
Accuracy	SVM : 77% CNN : 79% LR : 86%	SVM : 81% CNN : 90% LR : 82%	SVM : 87% CNN : 97% LR : 86%
Input Features	Pregnancies, glucose, blood pressure, insulin, bmi, Skin Thickness, Diabetes Pedigree Function etc.	Age, sex, chest pain type, cholesterol, resting blood pressure, fasting blood sugar etc	22 numerical features from the dataset like (MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%))
Evaluation Metrics	Accuracy, Precision, F1-score	Accuracy, Precision, F1-score	Accuracy, Precision, Recall, F1 Score.
Activation Function	Sigmoid, ReLU	ReLU, Softmax/Sigmoid	ReLU, Sigmoid
Optimizer	Adam	SGD	Adam
CNN layers	Input layer, Conv2D, batch normalisation, maxpooling2D, flatten layer, dense layer.	Input layer, Conv2D, batch normalisation, maxpooling2D, Dropout layer, output layer	Conv1D layers, BatchNormalization, Dropout, Flatten, Dense layers.
Frameworks & Library	TensorFlow, NumPy, Pandas, Scikit-learn.	TensorFlow, Keras, NumPy, Pandas, Scikit-learn	TensorFlow, Keras, NumPy, Pandas, Scikit-learn.
Terminology	ML, DL, Feature Scaling, Loss Function, Overfitting & Underfitting, Hyperparameter Tuning	ML & DL, Binary Classification Model Evaluation Metrics, Gradient Descent, Neural Network Layers	Conv1D, BatchNormalization, Dropout, Flatten, Dense Layer, Binary Crossentropy Loss

IV. CONCLUSION

Disease detection system focused on three major disease such as diabetes disease, heart disease and Parkinson's disease using machine learning and deep learning algorithms. Real-time patient data is use for train and test models. There are three algorithms use for model train and test such as Support Vector Machine (SVM), Logistic Regression (LR), and Convolutional Neural Network (CNN), allowing analysis of different approaches for each disease. The implementation of this model not just used for the improvements of prediction accuracy but it also show the computational method can support faster and more reliable medical diagnosis.

In diabetes prediction clinical data like glucose levels, BMI, age, insulin levels, and blood pressure were important in identifying at-risk patients. Logistic Regression achieved high performance 86% accuracy compare to Support vector machine and Convolutional Neural Network (CNN). In Parkinson's disease prediction Logistic Regression and SVM offered reliable results but Convolutional Neural Network perform higher performance than SVM and LR-logistic regression. It achieve 97% accuracy on test data which show its superior ability in processing voice-based biomedical features. In the heart disease prediction Convolutional Neural Network Model was select as better model compare to SVM and logistic regression. It achieve 90% accuracy although svm achieve 81% and logistic regression achieve 82% accuracy.

These research suggest that although traditional machine learning models are useful for quick, interpretable predictions and , deep learning approach offer more robust performance in cases of where data is high-dimensional or complex. In clinical decision support systems CNN serve as powerful tool with proper preprocessing and model tuning and helping in early diagnosis and improving patient outcomes.

The results in all three domains are promising that often achieve 90% accuracy in experimental settings, several challenges were faced. These contain issues like data quality, dataset imbalance, model interpretability, and clinical validation. Additionally its also concerns about patient privacy, bias in training data, and the need for transparent AI systems remain important for real-world deployment.

REFERENCES

[1] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Utilizing classification-based mining for diabetes diagnosis," *Int. J. Data Min. Knowl. Process.*, vol. 5, no. 1, pp. 1–14, 2015. doi: 10.5121/ijdkp.2015.5101.

[2] S. K. Sen and S. Dash, "Predictive modeling of diabetes using meta-level learning strategies," *J. Comput. Sci. Appl. Technol.*, vol. 2, no. 1, pp. 396–401, 2014..

[3] V. A. Kumari and R. Chitra, "Support vector machine classification for diabetes disease," *Int. J. Eng. Res. Appl.*, vol. 3, pp. 1797–1801, 2013.

[4] A. Sarwar and V. Sharma, "A Naive Bayes-based intelligent model for diagnosing type-2 diabetes," *Int. J. Comput. Appl.* (Special Issue: ICNICT), vol. 3, pp. 14–16, 2012. [Online]. Available: ijcaonline.org

[5] R. Arora and Suman, "Evaluation of classification methods across datasets using WEKA tool," *Int. J. Comput. Appl.*, vol. 54, no. 18, pp. 21–25, 2012. doi: 10.5120/8626-2492.

[6] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Approach for monitoring and diagnosing cardiovascular disease using computing models," *Int. J. Softw. Eng. Appl.*, vol. 9, no. 12, pp. 143–156, 2015.

[7] K. Vembandasamy, R. Sasipriya, and E. Deepa, "A probabilistic model for detecting cardiac disease using Naive Bayes," *Int. J. Innov. Sci. Eng. Technol.*, vol. 2, no. 3, pp. 441–444, 2015.

[8] V. Chaurasia and S. Pal, "Using data mining techniques to detect cardiovascular disease risks," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, no. 2, pp. 56–66, 2013.

[9] G. Parthiban and S. K. Srivatsa, "Machine learning applications for diagnosing cardiac conditions in diabetic patients," *Int. J. Appl. Inf. Syst.*, vol. 3, no. 7, pp. 25–30, 2012.

[10] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "Evolution-based method for selecting attributes in mining applications," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8616–8630, 2009, doi: 10.1016/j.eswa.2008.10.013.

[11] R. H. Abiyev and S. Abizade, "Fuzzy neural networks for Parkinson's diagnosis: An adaptive computing perspective," *Comput. Math. Methods Med.*, vol. 2016, Article ID 1267919, 2016. doi: 10.1155/2016/1267919

[12] T. V. Sriram, M. V. Rao, G. Narayana, and D. Kaladhar, "Analysis of Parkinson's using speech data and machine learning algorithms," in *Proc. 3rd Int. Conf. Frontiers Intell. Comput. Theory Appl. (FICTA)*, Springer, 2015, pp. 151–157.

[13] F. J. M. Shamrat et al., "Machine learning methods for Parkinson's disease prediction: A comparative analysis," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 2576–2580, 2019.

[14] S. Lahmiri and A. Shmuel, "Voice pattern ranking and optimized SVM for Parkinson's detection," *Biomed. Signal Process. Control*, vol. 49, pp. 427–433, 2019.

[15] Z. K. Senturk, "Machine learning-supported early stage diagnosis of Parkinson's disorder," *Med. Hypotheses*, vol. 138, p. 109603, 2020.

[16] H. Gunduz, "Classification of Parkinson's disease through deep neural analysis of speech data," *IEEE Access*, vol. 7, pp. 115540–115551, 2019.