MicroOhm Journal of Electrical and Computer Engineering
Vol. 1, No. 1; July 2025
E-ISSN: XXXX-XXXX

Implementation of Support Vector Machine (SVM) Method in Parkinson's Disease Classification Based on Acoustic Features

Bintang Hutagalung[1], Muhammad Nadjib Haeckal[2]

[1]Universitas Islam Negeri Sumatera Utara; bintanghutagalung121@gmail.com
[2]Universitas Indraprasta PGRI; nadjib.haeckal@gmail.com

ABSTRACT

Parkinson's disease is a progressive neurodegenerative disorder that significantly impacts quality of life and necessitates accurate early detection. Acoustic analysis of voice features offers a non-invasive and promising approach for classifying this condition. This study aims to evaluate the performance of the Support Vector Machine (SVM) algorithm in classifying Parkinson's disease based on 22 voice-related features extracted from a public dataset comprising 195 samples. The methodology includes data preprocessing (standardization and class weighting), model training using GridSearchCV, and evaluation based on standard classification metrics and diagnostic curves. The SVM model with an RBF kernel achieved an accuracy of 94.87%, precision of 96.55%, recall of 96.55%, F1-score of 96.55%, and a ROC-AUC score of 0.9828. The results indicate that SVM can effectively handle class imbalance and outliers without the need for complex techniques such as SMOTE or external feature selection. It is concluded that SVM is an effective method for early detection of Parkinson's disease based on voice data. Future research should focus on testing the model on larger and more diverse datasets and enhancing model interpretability for clinical use.

Keywords : *Parkinson's disease, Voice acoustics, Classification, Support vector machine, Machine learning;*

Corresponding Author:
Author Name: Bintang Hutagalung
Affiliation: Universitas Islam Negeri Sumatera Utara
Email: bintanghutagalung121@gmail.com

1. INTRODUCTION

Parkinson's disease is a progressive neurodegenerative disorder that affects both motor and non-motor systems (World Health Organization: WHO & World Health Organization: WHO, 2023). Based on 2021 Global Burden of Disease (GBD) data, the condition has affected 11.77 million individuals worldwide, reflecting a dramatic increase of 274% since 1990 (Li et al., 2025). Epidemiologic projections suggest that the number of sufferers will soar to 25.2 million by 2050, with global demographic shifts and population aging contributing approximately 89% of the increase (Ada.Enesco, 2025). In Indonesia, the estimated number of sufferers ranges from 200,000 to 400,000 people, with a projected increase from 90,000 cases in 2005 to 250,000 in 2030 (Su et al., 2025).

One of the main challenges in the management of Parkinson's disease lies in the complexity of accurate early diagnosis (Wasilewski et al., 2025). There is a time lag between the appearance of the first motor symptoms and the establishment of a definitive diagnosis, with a median time of 14.5 months (Cervantes-Arriaga et al., 2022). This delay is even more substantial in cases of young-onset Parkinson's disease (YOPD), which can range from 25 to 60 months. This phenomenon is due to the non-specific nature of early symptoms and overlap with other medical conditions (Riboldi et al., 2021). Prodromal symptoms such as olfactory disturbances (hyposmia) can appear up to 20 years before a motor diagnosis is made, while psychological manifestations such as anxiety or depression tend to appear one to two years earlier (Silva et al., 2022). Consequently, the clinical misdiagnosis rate reaches 15% to 24% when compared to post-mortem pathological confirmation (The Pathological Society of Great Britain & Ireland, 2024).

Given the limitations of currently available curative therapies, early detection is an important aspect of Parkinson's disease management. Early identification allows for more timely implementation of therapeutic interventions to slow symptom progression and optimize patient

quality of life. In this regard, the analysis of the acoustic characteristics of the voice of Parkinson's patients has shown substantial potential as a non-invasive biomarker (Gison et al., 2025). The scientific foundation of this approach is based on the high prevalence of vocal disorders (dysphonia) and motor speech disorders (hypokinetic dysarthria) that occur in approximately 89% of Parkinson's patients (Liu et al., 2025). These clinical manifestations are often one of the first indicators of motor dysfunction, as damage to the nigrostriatal pathway due to degeneration of dopaminergic neurons directly affects laryngeal muscle control (Lines et al., 2024).

Advances in computing technology have enabled the extraction and analysis of measurable acoustic features from voice recordings, including fundamental frequency variation (jitter), amplitude variation (shimmer), and harmonic-to-noise ratio (HNR). The integration of these acoustic features with machine learning algorithms has been shown to be effective in building classification models capable of distinguishing healthy individuals from Parkinson's sufferers with a high degree of accuracy. One of the previous studies used a combination of Particle Swarm Optimization (PSO) for feature selection and XGBoost for classification, and applied SMOTE to overcome class imbalance. The results showed that the AUC value of the model with feature selection without SMOTE and hyperparameter tuning was 0.9325, while the model without feature selection only reached 0.9250. When both SMOTE and hyperparameter tuning techniques were used together, the use of feature selection was able to provide a more substantial performance improvement, with the feature-selected model achieving an AUC value of 0.9483 compared to 0.9366 in the model without feature selection (Kurnia et al., 2023).

Although these results show promising performance, there is an opportunity to explore other algorithms that can provide high accuracy in the scope of Parkinson's disease classification. Support Vector Machine (SVM) offers unique characteristics that potentially provide advantages over ensemble approaches such as XGBoost (Surono et al., 2025) . SVMs have the fundamental ability to find the optimal separating hyperplane in a high-dimensional feature space through the concept of margin maximization, which enables superior generalization to never-before-seen data (Ginting et al., 2024) . The theoretical advantages of SVM lie in its ability to handle high feature dimensions without experiencing the curse of dimensionality, and the stability of its performance on datasets with relatively limited sample sizes (Pratama & Prasetyaningrum, 2025). SVM can also implicitly handle feature selection through kernel mechanism and regularization parameters, and can be adjusted to overcome class imbalance through proper class weight setting (Fajriyah et al., 2025).

Based on these potential advantages, this study proposes the implementation and comprehensive evaluation of a Support Vector Machine model for Parkinson's disease classification based on voice acoustic features. The main focus of this research is to answer fundamental questions regarding the performance of SVM models in classifying Parkinson's disease based on acoustic voice characteristics, and explore their effectiveness as an alternative to existing approaches. Through systematic evaluation using standard classification metrics and comprehensive diagnostic analysis, this research is expected to contribute to the understanding of SVM effectiveness in the domain of voice-based Parkinson's disease detection, while supporting the development of more accurate and non-invasive diagnostic tools for future clinical implementation.

2. LITERATURE REVIEW

1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It operates by finding the optimal hyperplane that separates data points from different classes in a high-dimensional space. In this study, SVM is employed to classify whether a person has Parkinson’s disease or not based on their voice acoustic features (Yahya, 2022).

2. Classification

Classification refers to the systematic arrangement of objects or data into groups or categories based on established rules or standards. It is the process of assigning labels to instances based on their characteristics. In this research, classification is used to predict whether an individual suffers from Parkinson’s disease based on their voice acoustic features (Sari, 2024).

3. Acoustic Features

Acoustic Features refers to the study of the physical properties of human speech sounds, including frequency, amplitude, and temporal characteristics. Changes in acoustic parameters, such as jitter and shimmer, are often associated with neurological disorders like Parkinson's disease, as they affect motor control of speech. In this study, voice acoustic features are extracted from patient recordings to be used as input for the Parkinson's disease classification model (Prasetyo, 2022).

4. Parkinson's Disease

Parkinson's disease is a progressive neurodegenerative disorder that affects the central nervous system, primarily impacting motor skills and speech. It is caused by the degeneration of nerve cells in the substantia nigra, a brain region responsible for producing dopamine. In this study, Parkinson's disease is the target of classification using voice acoustic features (Black & Hawks, 2022).

3. METHOD

The methodological approach of this study follows a structured workflow commonly used in machine learning research. It consists of five main stages: data collection, data preprocessing, model construction, model evaluation, and results analysis. Each stage is designed to ensure the integrity and validity of the classification process for Parkinson's disease based on acoustic voice features.

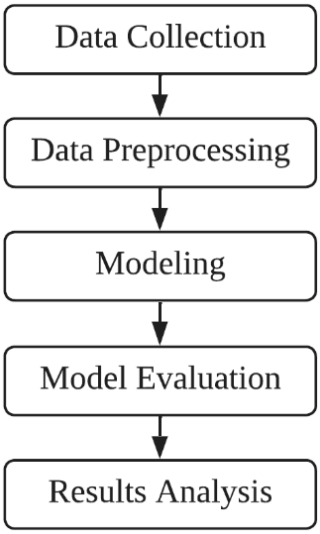


Figure 1. Flowchart

This study employs a quantitative method using the "Parkinson Disease Detection" dataset from Kaggle. The dataset contains 195 samples with 22 voice acoustic features and 1 binary target feature. Data preprocessing includes removal of the identifier feature, stratified splitting of the dataset into training (80%) and testing (20%) sets, and feature standardization using StandardScaler. The Support Vector Machine (SVM) classification model is built within a pipeline, where hyperparameter optimization (C, gamma, kernel) is performed using GridSearchCV with 5-fold cross-validation, utilizing accuracy as the scoring metric.

Class imbalance is addressed by setting class_weight='balanced'. Model evaluation on the test set is conducted using the following performance metrics:

1. Accuracy: measures the proportion of total correct predictions out of all predictions made.
   $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$
2. Precision: measures the proportion of correctly predicted positive instances compared to all positive predictions.
   $$Precision = \frac{TP}{TP+FP} \quad (2)$$
3. Recall: measures the model's ability to detect true positive instances.
   $$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-Score is the harmonic mean between precision and recall, provides a balance between the two.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

5. ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): measures the model's ability to distinguish between the positive and negative classes across various classification thresholds. The AUC score represents the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one.

4. RESULTS AND DISCUSSION

4.1 Results

The research dataset consists of 195 samples with a total of 24 features, which include 22 numeric acoustic voice features, one identification feature ('name'), and one target variable ('status') indicating the health condition, i.e. 0 for healthy, and 1 for Parkinson's.

Table 1. Example of Parkinson Disease Detection Dataset

Name	phon_R01_S01_1	phon_R01_S01_2	phon_R01_S01_3	phon_R01_S01_4
MDVP:Fo(Hz)	119.992	122.4	116.682	116.676
MDVP:Fhi(Hz)	157.302	148.65	131.111	137.871
MDVP:Flo(Hz)	74.997	113.819	111.555	111.366
MDVP:Jitter(%)	0.00784	0.00968	0.0105	0.00997
MDVP:Jitter(Abs)	0.00007	0.00008	0.00009	0.00009
MDVP:RAP	0.0037	0.00465	0.00544	0.00502
MDVP:PPQ	0.00554	0.00696	0.00781	0.00698
Jitter:DDP	0.01109	0.01394	0.01633	0.01505
MDVP:Shimmer	0.04374	0.06134	0.05233	0.05492
MDVP:Shimmer(dB)	0.426	0.626	0.482	0.517
Shimmer:APQ3	0.02182	0.03134	0.02757	0.02924
Shimmer:APQ5	0.0313	0.04518	0.03858	0.04005
MDVP:APQ	0.02971	0.04368	0.0359	0.03772
Shimmer:DDA	0.06545	0.09403	0.0827	0.08771
NHR	0.02211	0.01929	0.01309	0.01353
HNR	21.033	19.085	20.651	20.644
status	1	1	1	1
RPDE	0.414783	0.458359	0.429895	0.434969
DFA	0.815285	0.819521	0.825288	0.819235
spread1	-4.813031	-4.075192	-4.443179	-4.117501
spread2	0.266482	0.33559	0.311173	0.334147
D2	2.301442	2.486855	2.342259	2.405554
PPE	0.284654	0.368674	0.332634	0.368975

Descriptive statistics for numerical features, which present measures of central tendency (mean, median), dispersion (standard deviation, interquartile range), and minimum and maximum values.

Table 2 . Dataset Descriptive Statistics

	count	mean	std	min	25%	50%	75%
MDVP:Fo(Hz)	195.0	154.228641	41.390065	88.333000	117.572000	148.790000	182.769000
MDVP:Fhi(Hz)	195.0	197.104918	91.491548	102.145000	134.862500	175.829000	224.205500

MDVP:Flo(Hz)	195.0	116.324631	43.521413	65.476000	84.291000	104.315000	140.018500
MDVP:Jitter(%)	195.0	0.006220	0.004848	0.001680	0.003460	0.004940	0.007365
MDVP:Jitter(Abs)	195.0	0.000044	0.000035	0.000007	0.000020	0.000030	0.000060
MDVP:RAP	195.0	0.003306	0.002968	0.000680	0.001660	0.002500	0.003835
MDVP:PPQ	195.0	0.003446	0.002759	0.000920	0.001860	0.002690	0.003955
Jitter:DDP	195.0	0.009920	0.008903	0.002040	0.004985	0.007490	0.011505
MDVP:Shimmer	195.0	0.029709	0.018857	0.009540	0.016505	0.022970	0.037885
MDVP:Shimmer(dB)	195.0	0.282251	0.194877	0.085000	0.148500	0.221000	0.350000
Shimmer:APQ3	195.0	0.015664	0.010153	0.004550	0.008245	0.012790	0.020265
Shimmer:APQ5	195.0	0.017878	0.012024	0.005700	0.009580	0.013470	0.022380
MDVP:APQ	195.0	0.024081	0.016947	0.007190	0.013080	0.018260	0.029400
Shimmer:DDA	195.0	0.046993	0.030459	0.013640	0.024735	0.038360	0.060795
NHR	195.0	0.024847	0.040418	0.000650	0.005925	0.011660	0.025640
HNR	195.0	21.885974	4.425764	8.441000	19.198000	22.085000	25.075500
status	195.0	0.753846	0.431878	0.000000	1.000000	1.000000	1.000000
RPDE	195.0	0.498536	0.103942	0.256570	0.421306	0.495954	0.587562
DFA	195.0	0.718099	0.055336	0.574282	0.674758	0.722254	0.761881
spread1	195.0	-5.684397	1.090208	-7.964984	-6.450096	-5.720868	-5.046192
spread2	195.0	0.226510	0.083406	0.006274	0.174351	0.218885	0.279234
D2	195.0	2.381826	0.382799	1.423287	2.099125	2.361532	2.636456
PPE	195.0	0.206552	0.090119	0.044539	0.137451	0.194052	0.252980

Analysis of the target class distribution revealed an imbalance, with 147 samples (75.4%) in the Parkinson's class (status 1) and 48 samples (24.6%) in the healthy class (status 0), resulting in a minority to majority class ratio of 0.33.

Varying degrees of skewness were observed in some of the numerical features; for example, the NHR feature showed high positive skewness (4.22), while the HNR feature showed negative skewness (-0.51). Detail of skewness values. Identification of outliers using the Interquartile Range (IQR) method with a factor of 1.5*IQR indicated the presence of outliers in a number of features, for example, the NHR feature had 19 outliers and MDVP:Fhi(Hz) had 11 outliers (Table 3). The identified outliers were not specifically addressed (e.g., removed or transformed) to maintain the integrity of the original data, but rather, their influence was mitigated through the use of SVM algorithms that are relatively robust to outliers and through the feature standardization process.

Table 3. Skewness and Number of Outliers (IQR Method)

Name	Skewness	Outliers
MDVP:Fo(Hz)	0.591737	0
MDVP:Fhi(Hz)	2.542146	11
MDVP:Flo(Hz)	1.217350	9
MDVP:Jitter(%)	3.084946	14
MDVP:Jitter(Abs)	2.649071	7
MDVP:RAP	3.360708	14
MDVP:PPQ	3.073892	15
Jitter:DDP	3.362058	14
MDVP:Shimmer	1.666480	8
MDVP:Shimmer(dB)	1.999389	10
Shimmer:APQ3	1.580576	6
Shimmer:APQ5	1.798697	13
MDVP:APQ	2.618047	12
Shimmer:DDA	1.580618	6
NHR	4.220709	19
HNR	-0.514317	13
status	-1.187727	-
RPDE	-0.143402	0
DFA	-0.033214	0
spread1	0.432139	4
spread2	0.144430	2
D2	0.430384	1
PPE	0.797491	5

The data pre-processing stage begins with splitting the data into feature sets (X), after removal of identification features ('name'), and target variables (y). The data was further divided into training (80%, 156 samples) and testing (20%, 39 samples) sets using a stratification method based on the target class with randomization (random_state=42). This stratification resulted in similar class proportions in both data sets: the training set consisted of 118 class 1 samples and 38 class 0 samples, while the testing set consisted of 29 class 1 samples and 10 class 0 samples. The numerical features in both sets were then standardized using StandardScaler. This standardization process scales each feature so that it has a mean close to zero and a standard deviation close to one in the training set.

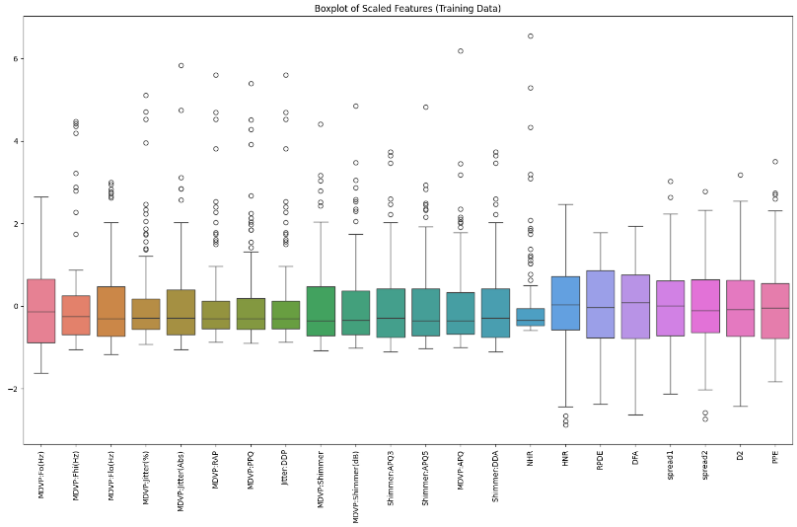


Figure 1. Boxplot of Standardized Features (Training Data)

Based on the figure (Figure 2), the statistical distribution of each numerical feature in the training set after the standardization process. All the features have medians around zero, indicating

that the standardization was successful. The interquartile range (IQR) is relatively uniform across most of the features, however, some features such as *MDVP:Fhi(Hz)*, *Jitter (%)*, *Shimmer*, *NHR*, and *PPE* display a considerable number of outliers, indicated by dots outside the whiskers. The presence of these outliers indicates an abnormal or skewed distribution of data on certain features. On the other hand, features such as *DFA*, *RPDE*, and *spread1* show a more symmetrical and compact distribution. Correlations between features on the standardized training data, which showed several feature pairs with moderate to high correlations (e.g., between MDVP:Jitter(%) and MDVP:Jitter(Abs) with r > 0.9).

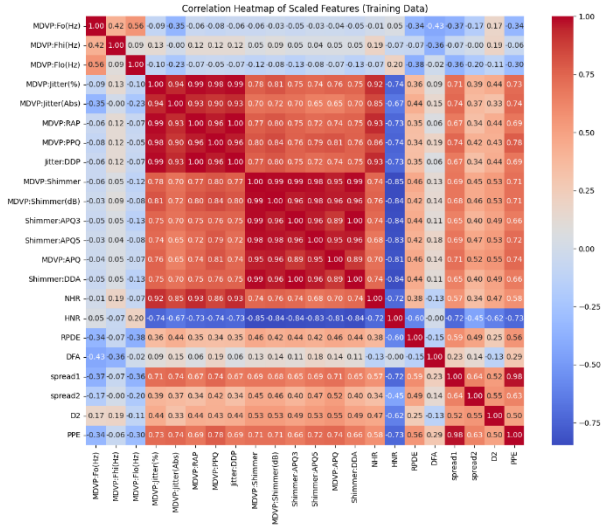


Figure 2. Heatmap of Standardized Feature Correlation (Training Data)

The representation of the training data in two principal components using Principal Component Analysis (PCA), which accounts for X% of the cumulative variance, indicates a visual separation between classes despite the overlap.

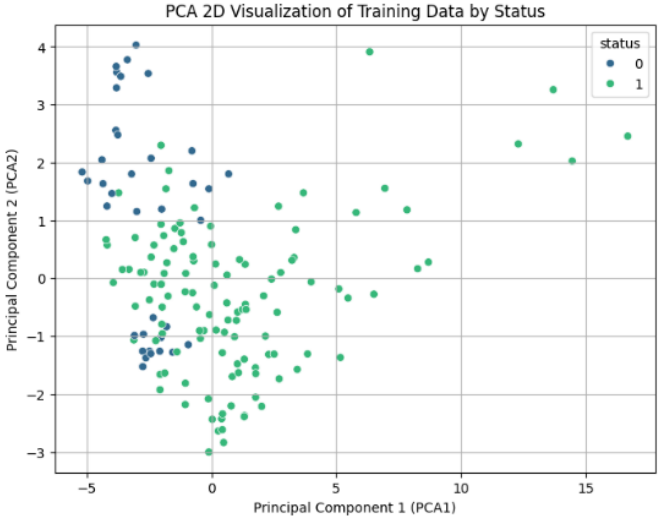


Figure 3. 2D PCA Visualization of Training Data by Status

Support Vector Machine (SVM) classification models are developed using a pipeline structure that integrates the feature standardization stage with the SVM model. This pipeline approach ensures consistency of data pre-processing applications during the training and evaluation phases. SVM model hyperparameter optimization was performed via GridSearchCV with a 5-fold cross-validation strategy on the training dataset. The parameter space explored included regularization parameter C ([0.1, 1, 10, 100]), gamma kernel coefficients (['scale', 'auto', 0.1, 1]), and SVM kernel type (['linear', 'rbf', 'poly', 'sigmoid']), resulting in a total of 64 tested parameter combinations. The results of the hyperparameter optimization process showed that the best parameter combination for the SVM model was C=100, gamma=0.1, and RBF kernel. The model with this configuration achieved an

average cross-validation accuracy of 0.9421 on the training dataset, with inter-fold accuracy scores ranging from 0.9032 to 0.9688. The difference between the highest training score on a particular fold and the lowest validation score on another fold does not indicate any significant overfitting at this stage.

The optimal model generated by GridSearchCV was further evaluated using the test data set. The classification accuracy of the model on the test data was recorded as 0.9487, which means 37 out of 39 test samples were correctly classified. For the positive class (status 1, Parkinson's), the precision metric was 0.9655 (28 TP/(28 TP + 1 FP)), recall was 0.9655 (28 TP/(28 TP + 1 FN)), and F1-score was 0.9655. For the negative class (status 0, healthy), the precision is 0.9000 (9 TN / (9 TN + 1 FN)), the recall is 0.9000 (9 TN / (9 TN + 1 FP)), and the F1-score is 0.9000. These metrics were generated using the classification_report function of scikit-learn. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) on the test data was 0.9828, which indicates the model's strong discrimination ability between the two classes. As a baseline, the accuracy of the dummy classifier that always predicts the majority class (state 1) on the test data is 29/39 ≈ 0.7436.

The types of errors made by the model are evaluated in more detail through the confusion matrix (Figure 5), which maps the relationship between predictions and actual labels. The confusion matrix shows 28 True Positives (TP), 9 True Negatives (TN), 1 False Positive (FP), and 1 False Negative (FN). These values indicate that the model was able to correctly classify most of the samples in both the positive (Parkinson's) and negative (healthy) classes. The very low error ratio of only one error in each type (FP and FN) indicates that the model has a good balance of performance in detecting the presence or absence of disease. However, False Negative errors remain a major concern in the medical context, as one Parkinson's patient was not identified by the model. Such errors have the potential to delay diagnosis and treatment, so in real-world applications additional approaches should be considered to reduce the risk of FN without increasing FP.

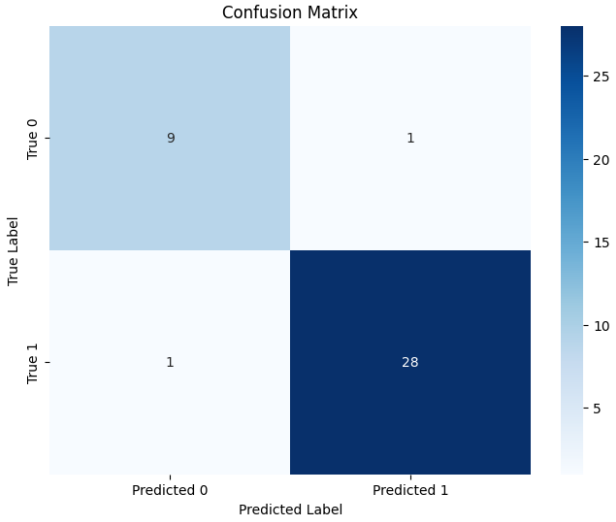


Figure 4. Confusion Matrix

A comprehensive classification report, including precision, recall, F1-score, and support (actual number of samples per class: 10 for class 0, 29 for class 1) for each class.

Classification Report:

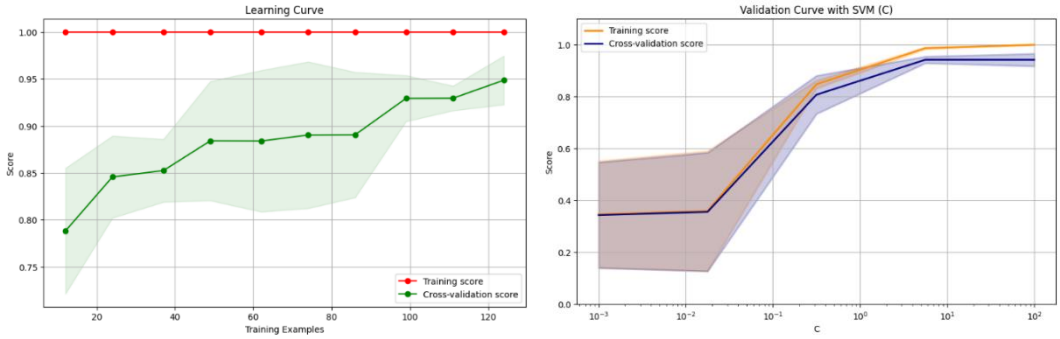
	precision	recall	f1-score	support
0	0.90	0.90	0.90	10
1	0.97	0.97	0.97	29
accuracy			0.95	39
macro avg	0.93	0.93	0.93	39
weighted avg	0.95	0.95	0.95	39

Figure 5. Classification Report

To verify the stability of the model performance, an additional evaluation was performed using 5-fold cross-validation on the entire training data, using the same pipeline (including standardization

applied per fold to avoid data leakage). The resulting cross-validation accuracy scores at each fold were [0.96875, 0.96774194, 0.93548387, 0.90322581, 0.93548387], with a mean value of 0.9421 and a standard deviation of 0.0244. These results are consistent with the best scores reported by GridSearchCV, indicating the stability of the model performance on the training data.

The learning curves generated by varying the training sample size from 10% to 100% (10 data points), illustrate the evolution of the training and cross-validation scores. It is observed that both curves converge at a high level of performance (final average training score ~0.99, final average validation score ~0.94), with a relatively small gap between training and validation scores at larger sample sizes. This pattern indicates that the model has good generalization and does not suffer from significant overfitting, despite the slightly higher training score (Figure 7, (a)). In the tuning analysis, validation curves for the parameter C (with a range of values [0.001, 0.01, 0.1, 1, 10] and gamma and kernel fixed at optimal values) were used to evaluate the sensitivity of the model to variations in the value of this parameter. The curves show that the model performance increases as the value of C increases until it reaches a plateau at C=10, and slightly decreases at C=100 for the validation score, confirming that the range used in GridSearchCV is sufficient.



(a) Learning Curve **(b)** Validation Curve

Figure 6. (a) Learning Curve; **(b)** Validation Curve

The ROC curve visualizes the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) at various classification threshold values. Based on Figure 10, the ROC curve generated by the model has a concave shape to the upper left and away from the reference diagonal line, which indicates excellent classification performance. The area under the curve (AUC) value of 0.98 indicates the model's very high discriminative ability in distinguishing between the positive (Parkinson's) and negative (healthy) classes. The initial points on the curve show that the model can achieve a TPR above 0.85 even at very low FPR levels (below 0.15). This visualization confirms that the model is not only accurate in general, but also reliable in maintaining sensitivity to positive cases without sacrificing much in mispredicting negatives as positives.

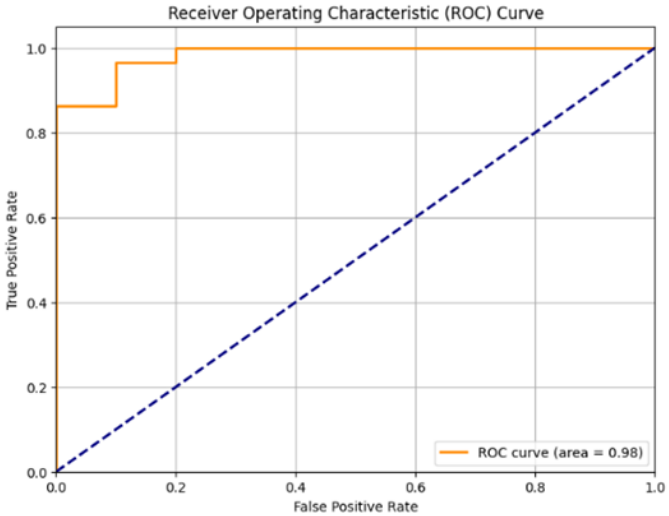


Figure 7. Receiver Operating Characteristic (ROC) Curve

4.2 Discussion

This study aims to evaluate the performance of the Support Vector Machine (SVM) model in classifying Parkinson's disease based on acoustic voice features. The results obtained show that the developed SVM model has good classification ability. The optimal model, which utilizes the RBF kernel with parameters C=100 and gamma=0.1, achieves an accuracy of 0.9487 on the test dataset, which substantially surpasses the baseline accuracy of the dummy classifier (0.7436). The ROC-AUC value of 0.9828 indicates a very strong discrimination ability between the healthy and Parkinson's classes. The balanced precision, recall, and F1-score metrics for the positive (Parkinson's) class, at 0.9655, and the moderately good metrics for the negative (healthy) class with precision, recall, and F1-score of 0.9000, indicate that the model is not only accurate overall, but, effective in identifying both classes.

The target class distribution in the dataset shows an imbalance, with the number of samples with Parkinson's disease (state 1) being more dominant than healthy samples (state 0). However, the use of the class_weight='balanced' parameter in the SVM model proved to be quite effective in addressing this issue, as reflected by the relatively high recall metrics for both classes. This class weighting approach tends to be more stable than oversampling techniques such as SMOTE, which may increase the risk of overfitting, especially on datasets of limited size such as the one used in this study. The confusion matrix on the test data (Figure 5) shows that the model committed only one False Positive and one False Negative misclassification. The False Negative error, which classifies individuals with Parkinson's as healthy, has more significant clinical implications as it can delay early intervention and therefore demands special attention if the model is adopted in a real medical setting.

Prior to the modeling stage, several data pre-processing steps were performed. The process of feature standardization using StandardScaler is an important step, given the sensitivity of the SVM algorithm to feature scale. Visualization of the standardized feature boxplots (Figure 2) confirmed that the features were comparable in scale. Correlation analysis (Figure 3) shows that there are some feature pairs with moderate to high correlation, however, SVM with RBF kernel generally handles multicollinearity well. Outlier identification (Table 3) shows the presence of extreme values in some features. In this study, no transformation was performed to address skewness or explicit outlier handling as these approaches prioritized keeping the data close to the original distribution and testing the ability of SVMs under these conditions. The good performance of the model indicates that the SVM algorithm with the right parameters is able to overcome the influence of the outliers, or that the outliers may be a valid variation of the data within the scope of Parkinson's voice features.

The learning curve (Figure 7 (a)) provides important information about the generalization of the model. The convergence between training scores and cross-validation scores at a high level of performance, with validation scores consistently above 0.94, indicates that the model does not suffer from significant overfitting and has good generalization ability to new data. The small difference between the final average training score (~0.99) and the final average validation score (~0.94) is still within reasonable limits for a highly capable non-linear model such as SVM with RBF kernel. The validation curve for the parameter C (Figure 7 (b)) confirms that the parameter range explored in GridSearchCV is adequate, with the model performance reaching a plateau at a certain value of C.

Compared to previous research by Kurnia et al. (2023) who used XGBoost with Particle Swarm Optimization (PSO) feature selection and class imbalance handling using SMOTE, this study shows competitive results. Kurnia et al. (2023) reported the highest AUC value of 0.9483 for the XGBoost model with feature selection, SMOTE, and hyperparameter tuning. The SVM model in this study, without explicit feature selection (other than that inherent in the SVM mechanism) and without oversampling techniques such as SMOTE (using class weighting only), achieved an ROC-AUC of 0.9828. This difference may be due to the characteristics of the dataset used, the difference in classification algorithms (SVM vs. XGBoost), or the effectiveness of SVM in handling high-dimensional data and class imbalance with weighting. SVMs, especially with RBF kernels, are known to find the optimal separating hyperplane in high-dimensional feature spaces. The absence of SMOTE or external feature selection in this study can be seen as a strength in terms of model simplicity and potential robustness to data variation, however, it can also be a weakness if irrelevant features or noise dominate in other datasets.

The results show that an appropriately configured SVM model is able to provide high classification performance for Parkinson's disease based on acoustic voice features, even on unbalanced datasets and without complex external feature selection. This contributes to the understanding of the effectiveness of SVM in this application domain and answers the research problem.

5. CONCLUSION

The Support Vector Machine (SVM) model developed in this study shows excellent classification performance in distinguishing Parkinson's sufferers from healthy individuals based on acoustic voice features. The best model with RBF kernel, C=100, and gamma=0.1 resulted in an accuracy of 94.87%, with F1-score, precision, and recall values for the positive class of 0.9655, respectively, and precision, recall, and F1-score values for the negative class of 0.9000. The confusion matrix showed only one error in each type (1 False Positive and 1 False Negative), and the ROC-AUC value of 0.98 confirmed the high discriminative ability of the model. These results support the potential utilization of SVM as a non-invasive method in voice-based Parkinson's disease early detection.

Although the evaluation results show good performance, the external validity of the model needs to be strengthened through testing on larger, varied and representative datasets from real populations. Future research is recommended to explore feature selection techniques, more sophisticated class imbalance handling strategies, and model interpretability approaches such as SHAP or LIME. It is important that the models are not only accurate, but also understandable and trustworthy in clinical contexts that demand transparency in decision-making.

REFRENCES

Ada.Enesco. 202AD. “Parkinson’s Disease Cases Predicted to Double by 2050: Global Study - EMJ.” EMJ. Retrieved June 22, 2025 (https://www.emjreviews.com/general-healthcare/news/parkinsons-disease-cases-predicted-to-double-by-2050-global-study/).

Black, M. oyce, and Jane Hokanson Hawks. 2022. “KMB: Gangguan Sistem Neurologis.” Elsevier Health Sciences 1–325.

Cervantes-Arriaga, Amin, Cynthia Sarabia-Tapia, Oscar Esquivel-Zapata, Susana López-Alamillo, Etienne Reséndiz-Henriquez, Teresa Corona, and Mayela Rodríguez-Violante. 2022. “Pitfalls and Caveats in the Diagnostic Pathway of People with ParkinsonÓ?S Disease.” Revista Mexicana de Neurociencia 23(4):119–25. doi: 10.24875/rmn.21000041.

Fajriyah, Nurul, Nouval Trezandy Lapatta, Deny Wiria Nugraha, and Rahmah Laila. 2025. “IMPLEMENTASI SVM DAN SMOTE PADA ANALISIS SENTIMEN MEDIA SOSIAL X TERHADAP PELANTIKAN AGUS HARIMURTI YUDHOYONO.” JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika) 10(2):1359–70.

Ginting, Joel Arie Putranta, Radiatun Maya Sari, Muhammad Rafli Dewantara Siregar, and Dedi Kiswanto. 2024. “ANALISIS SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI JENIS KELAMIN PADA IKAN CUPANG DENGANBANTUAN LOCAL BINARY PATTERN (LBP).” JATI (Jurnal Mahasiswa Teknik Informatika) 8(6):12782–86.

Gison, Annalisa, Marco Ruggiero, Davide Tufarelli, Stefania Proietti, Daniela Moscariello, and Marianna Valente. 2025. “Intensive Speech Therapy for Hypokinetic Dysarthria in Parkinson’s Disease: Targeting the Five Subsystems of Speech Production with Clinical and Instrumental Evaluation.” NeuroSci 6(1). doi: 10.3390/neurosci6010007.

Kurnia, Deni, Muhammad Itqan Mazdadi, Dwi Kartini, Radityo Adi Nugroho, and Friska Abadi. 2023. “Seleksi Fitur Dengan Particle Swarm Optimization Pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost.” Jurnal Teknologi Informasi Dan Ilmu Komputer 10(5):1083–94. doi: 10.25126/jtiik.20231057252.

Leite Silva, Ana Beatriz Ramalho, Roger Wilson Gonçalves de Oliveira, Guilherme Pinheiro Diógenes, Marina Feitosa de Castro Aguiar, Camilla Costa Sallem, Micael Porto Portela Lima,

Luciano Barroso de Albuquerque Filho, Sara Diógenes Peixoto de Medeiros, Lucas Lopes Penido de Mendonça, Paulo Cesar de Santiago Filho, Diogo Pasquali Nones, Pamella Mendes Martiniano da Silva Cardoso, Michelle Zonkowski Ribas, Stéfani Lara Galvão, Gabriel Felipe Gomes, Amanda Rebouças Bezerra de Menezes, Nayla Lima Dos Santos, Victor Monteiro Mororó, Fairane Sousa Duarte, and Júlio César Claudino Dos Santos. 2023. "Premotor, Nonmotor and Motor Symptoms of Parkinson's Disease: A New Clinical State of the Art." Ageing Research Reviews 84:101834. doi: 10.1016/j.arr.2022.101834.

Li, Mimi, Xiaofang Ye, Zhengping Huang, Lichao Ye, and Chunnuan Chen. 2025. "Global Burden of Parkinson ' s Disease Based from 1990 to 2021 : A Population- ¬ Study." BMJ Open 1–10. doi: 10.1136/bmjopen-2024-095610.

Liu, Vicki, Dori Smith, and Helena Yip. 2025. "Prevalence and Treatment of Dysphonia in Parkinson's Disease: A Cross-Sectional National Database Study." Laryngoscope Investigative Otolaryngology 10(3):e70149. doi: 10.1002/lio2.70149.

Meeting, Winter. 2024. "WINTER MEETING Pathology and Precision Medicine 30 – 31." The Pathological Society of Great Britain and Ireland (January):1–38. doi: 10.1002/path.6337.

OPEN LINES. 2024. "How Does Parkinson's Disease Affect Speech? | Open Lines®." OPEN LINES.

Prasetyo, Aldy Dwi. 2022. "Lebih Dekat Dengan Aeroakustik: Closer to Aeroacoustics." Bugocenter 1–23.

Pratama, Harfin Ibna, and Putri Taqwa Prasetyaningrum. 2025. "Penerapan Support Vector Machine Untuk Analisis Sentimen Pada Google Review Hotel." Journal of Information System Research (JOSH) 6(2):1244–52. doi: 10.47065/josh.v6i2.6645.

Riboldi, Giulietta M., Emanuele Frattini, Edoardo Monfrini, Steven J. Frucht, and Alessio Di Fonzo. 2022. "A Practical Approach to Early-Onset Parkinsonism." Journal of Parkinson's Disease 12(1):1–26. doi: 10.3233/JPD-212815.

Sari, Rahayu Mayang. 2024. Klasifikasi Data Mining. Serasi Media Teknologi.

Su, Dongning, Yusha Cui, Chengzhang He, Peng Yin, Ruhai Bai, Jinqiao Zhu, Joyce S. T. Lam, Junjiao Zhang, Rui Yan, Xiaoqing Zheng, Jiayi Wu, Dong Zhao, Anxin Wang, Maigeng Zhou, and Tao Feng. 2025. "Projections for Prevalence of Parkinson's Disease and Its Driving Factors in 195 Countries and Territories to 2050: Modelling Study of Global Burden of Disease Study 2021." BMJ (Clinical Research Ed.) 388:e080952. doi: 10.1136/bmj-2024-080952.

Surono, Muhammad, Muhammad Fadli, Dian Sri Purwamti, Erliyan Redy Susanto, Magister Ilmu Komputer, Universitas Teknokrat Indonesia, Kota Bandar Lampung, Politeknik Negeri Lampung, and Kota Bandar Lampung. 2025. "Hybrid XGBoost-SVM Model Untuk Sistem Pendukung Keputusan Dalam Prediksi Penyakit Diabetes." INSOLOGI: Jurnal Sains Dan Teknologi 4(3):443–54. doi: 10.55123/insologi.v4i3.5410.

Wasilewski, Andrzej, Eliza Wasilewska, and Agata Serrafi. 2025. "Exploring Diagnostic Markers and Therapeutic Targets in Parkinson's Disease: A Comprehensive (1)H-NMR Metabolomic Analysis - Systematic Review." Archivum Immunologiae et Therapiae Experimentalis 73(1). doi: 10.2478/aite-2025-0011.

WHO. 2023. "Parkinson Disease." World Health Organization. Retrieved June 22, 2025 (https://www.who.int/news-room/fact-sheets/detail/parkinson-disease).

Yahya. 2022. Data Mining. CV Jejak (Jejak Publisher).