Machine Learning-Based Approach to Improving Parkinson's Disease Diagnosis Through Voice Signal Analysis.

by

Arman Hossain Shanto
18301255
Upoma Deb Suchi
20201109
Raida Jafar Chowdhury
20301026
Afnan Mohammed
21101005
Fairuz Suhala Reza
22141029

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Arman Hossain Shanto
18301255

Upoma Deb Suchi
20201109

Raida Jafar Chowdhury
20301026

Afnan Mohammed
21101005

Fairuz Suhala Reza
22141029

Approval

The thesis/project titled "Machine Learning-Based Approach to Improving Parkinson's Disease Diagnosis Through Voice Signal Analysis"

1. Arman Hossain Shanto (18301255)
2. Upoma Deb Suchi (20201109)
3. Raida Jafar Chowdhury (20301026)
4. Afnan Mohammed (21101005)
5. Fairuz Suhala Reza (22141029)

Of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October, 2025.

Examining Committee:

Supervisor:
(Member)

Md Tanzim Reza
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Parkinson's disease (PD) is a chronic neurodegenerative disorder that influences motor and non-motor function and imposes huge health-related costs on society. Early and precise diagnosis still presents a big challenge because clinical examinations are subjective, and diagnostic methods based on imaging are expensive and invasive. As the vocal deficiencies are usually present in early PD, voice is considered to be a potential non-invasive and widely available neurobiological marker for early diagnosis.This thesis presents an end-to-end framework for PD detection via ML and XAI. Voice datasets were preprocessed for normalization, class balancing and dimensionality reduction in order to improve data quality. Acoustic Feature Selection We used a range of feature selection methods, including statistical significance testing, SelectKBest ranking, and principal component analysis to extract discriminative sub-sets of acoustic features. Various machine learning models were trained and tested, the ensemble methods performed well. The results indicate that the oversight model based on the selection of only 100 picked features with XGBoost approach provides high accuracy, precision and recall as well as F1-scores greater than 0.85 on the test set. SHAP and Morris Sensitivity Analysis were used to increase the interpretablity, revealing important indicators of PD-related vocal changes including jitter, shimmer and MFCC features. These observations were consistent with clinical data, which linked computational modeling to biological knowledge. The results indicate that voice coupled with interpretable ML techniques can be used as a robust digital biomarker for PD. This work will help build scalable, accessible and explainable tools for early diagnosis and remote healthcare.

Keywords: Parkinson's disease, Voice analysis, Machine learning, Ensemble learning, XGBoost, Acoustic features, Explainable AI, SHAP, Morris Sensitivity Analysis, Digital biomarkers

Acknowledgement

Firstly, all praise and gratitude to Almighty Allah, whose blessings have allowed us to complete this thesis successfully and without major interruption.
We would like to express our sincere thanks to our co-advisor, Md Tanzim Reza, for his constant guidance, valuable advice, and kind support throughout this work. His help was invaluable whenever we encountered difficulties.
Finally, we would like to extend our heartfelt thanks to our parents.Without their unwavering support, encouragement, and prayers, reaching this stage of our graduation would not have been possible. Their guidance and care have been a constant source of motivation throughout our academic journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background on Parkinson's Disease and Voice-Based Detection

Parkinson disease (PD) is one of the most common neurodegenerative diseases with a current estimated worldwide prevalence of more than 10 million people. It is mainly due to the gradual degeneration of dopaminergic neurons in the substantia nigra pars compacta, a key midbrain area that mediates motor control. The loss of dopamine interferes with the communication between the basal ganglia and motor cortex and contributes to classic features such as resting tremor, bradykinesia, rigidity, and postural instability. Non-motor features, such as depression, anxiety, cognitive impairment, sleep disturbances and autonomic dysfunction become increasingly prominent with the progression of disease; consequently they further decrease patients' quality of life [10][14].

PD imposes a great socioeconomic burden. Direct costs are constituted by hospitalizations, medications and therapy; indirect costs include lost productivity, early retirement and the need of long-term care. In the USA, it is estimated to cost almost 1 million dollars per patient annually throughout their life. With increasing life expectancy worldwide, the PD population is set to rise and early detection and intervention are essential [13].

The early diagnosis of PD is still difficult. Currently, clinicians SBs largely depend on the Unified Parkinson's Disease Rating Scale (UPDRS) as well as clinical examinations to quantify the symptoms; however, they are subjective and subject to variability based on physician experience. Neuroimaging techniques including dopamine transporter single-photon emission computed tomography (DaT-SPECT) and positron emission tomography (PET), although objective, are expensive, invasive and tend to diagnose PD at more of an advanced stage when neuronal loss is already substantial. This delay between 2 and 5 years from the first onset of symptoms and diagnosis is a lost opportunity for disease modifying interventions [2].

In recent years, the human voice has been proposed as a potential biomarker for early and non-invasive diagnosis of PD. Speech problems is represented by a group of close to 90% of the patients at some stage of the disease, termed hypokinetic dysarthria (HD). It presents as decreased vocal volume or loudness (hypophonia), monotonic pitch, imprecise articulation, breathiness, and shorter duration of speech. Crucially, these vocal alterations frequently precede the development of typical mo-

tor symptoms, providing a potential window for earlier diagnosis [4].
Rather, voice and speech analysis provides several advantages compared to standard diagnostic approaches. It is inexpensive, non-invasive and remote (i.e. could be carried out using off-the-shelf equipment such as smartphones). Due to the use of acoustic signal processing and machine learning methodologies, mild patterns of dysphonia can be identified and used to distinguish PD patients from normal subjects. Among these features are jitter, shimmer, harmonics-to-noise ratio (HNR), Mel-Frequency Cepstral Coefficients (MFCCs) and nonlinear dynamic measures such as Recurrence Period Density Entropy (RPDE) and Pitch Period Entropy (PPE). These acoustic biomarkers detect disturbances in vocal fold oscillation, resonance and prosody that have been tightly linked to the pathology of PD.
The rapid development of artificial intelligence (AI) and machine learning (ML) has taken PD diagnosis through voice detection studies into a higher gear. Making use of high-dimensional feature representations created from raw audio signals, ML models can learn to predict patterns human listeners cannot decipher. Voice analysis for feedback has therefore emerged as a thrusting frontier in computational neurology and digital health.

1.2 Problem Statement

However, despite the promising nature of voice as an effective biomarker for PD, a number of difficulties are Cambell to bringing it into clinical practice. First, data availability remains limited. Publicly available datasets such as UCI Parkinson's dataset have limited number of samples (only 195 recordings from just 31 individuals), hence end up training models that are not robust nor generalizable. Furthermore, large-scale publicly available datasets are often acquired in a controlled laboratory environment and do not quite capture the variety of natural conditions covering background noises, different micro- phones or dialects.
Another issue is the imbalance of classes. As the incidence of PD is relatively low, datasets tend to present more recordings from healthy than pathological subjects. This imbalance may cause models to be biased towards majority-class prediction that achieves high accuracy but poor sensitivity in predicting PD cases. Missed early cases are particularly worrisome as these patients can benefit most from prompt treatment.
Feature engineering presents further complexity. Speech signals are inherently high dimensional, and the extraction of acoustic, spectral and nolinear features result in several hundreds variables. Not all features are equally important for classification and redundant or irrelevant features can be noisy. While such a problem could be mitigated by a dimensionality reduction technique like PCA, it risks throwing away subtle though clinically relevant features. That means feature selection and optimization are indeed critical on creating efficient models.
Model development also poses challenges. Although conventional machine learning methods like Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (i.e., XGBoost [2], CatBoost [8]) have yielded acceptable performances they usually need well hyper-parameter tuning, and could fail to model the temporal relevance inspeech. Deep learning methods, like Convolutional Neural Network (CNN) [8] and Recurrent Neural Network (RNN) [9], excell in modeling deep features but require huge amounts of training data and computing resources. Moreover,

these models are criticized for being "black boxes," which may invoke skepticism on their interpretability and clinical reliability.
Finally, model calibration and explainability continue to be crucial problems. This is a common issue for algorithms which tend not to produce well-calibrated probability estimates, thus limiting their use in the clinic. Furthermore, clinicians need to understand the features influencing model predictions. XAI such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) provide insight of the model behavior, but are barely integrated in PD research.
In light of this exigency, a holistic framework that combines state-of-the-art machine learning algorithms, deep feature engineering, learning on balanced data, coupled with an explainability study is required to ensure accurate and interpretable voice-based PD detection.

1.3 Progress of Machine Learning in PD Voice Classification

The development of machine learning for voice PD detection mirrors trends in artificial intelligence as a whole. Statistical methods were prominent in early studies. Acoustic measures, including jitter and shimmer were extracted and compared (between patients and controls by t-tests/ANOVA). Although these approaches brought useful insights, the methods were limited both in terms of accuracy and scalability, commonly obtaining classification rates around 60–70% [].
Super-vised machine learning in the 2000s (Ventura et al. With datasets like the UCI Parkinson's voice dataset, scientists leveraged algorithms such as logistic regression, Naïve Bayes or SVMs. Particularly, SVMs were capable of modelling the non-linear boundaries to obtain accuracies superior to 85%. K-Nearest Neighbors (k-NN) classifiers were also widely used, as they based their decisions on similarities between neighborhoods and were directly interpretable.
Ensemble methods rise to prominence Ensemble methods came of age in the 2010s. Random Forests, an aggregation of many decision trees using bagging, emerged as a practical baseline model with better stability and interpretability. Randomness was increased with Extra Trees, in the sense of having lower variance and training time. The introduction of boosting methods, to name a few AdaBoost, Gradient Boosting as well as more recent XGBoost, CatBoost and LightGBM, increased classification accuracy above 95%. With respect to imbalanced datasets, feature selection and large sets of features these algorithms outperformed. Especially XGBoost was one of the most popular algorithms under the hood for PD voice classification since it performed better and could rank features importance.
Simultaneously, dimensionality reduction techniques evolved. The curse of dimensionality was overcome with techniques like RFE, PCA, and more complex nonlinear methods including t-SNE and autoencoders. These methods not only enhanced the effectiveness of models but also contributed to interpretability by discovering some important acoustic markers.
The past decade has witnessed the burgeoning of deep learning. CNNs have also been used as the front end for converting speech signals into spectrograms; in this way, discriminative features could be learned end-to-end. Temporal dependencies

within sequences of rhotic segmentations can be modelled with recurrent architectures (e.g., LSTMs, GRUs), and in doing so yield improved performance compared to classical systems on connected speech. Models that combine the two, using RNNs in conjunction with CNNs, have taken the field even further. Transformers 1 have also been used for self-supervised learning of representations from large audio corpora using Wav2Vec2 and HuBERT, and to enable transfer between PD detection tasks.

Explainable AI has been increasingly recognized as a vital companion to these advancements. Approaches such as SHAP and LIME are now gaining popularity for explaining feature contributions, indicating acoustic features as most relevant predictors of PD including shimmer, jitter, and MFCC coefficients. In addition to promoting trust in the clinician community, this interpretability leads model's decisions to be coherent with biomedical knowledge.

Despite these advances, gaps remain. A significant limitation is the use of small, monolingual datasets in many studies. Although deep learning methods are powerful, they tend to overfit in data-poor situation. In addition, to my knowledge research in this direction generally considers only accuracy but fails to investigate calibration, robustness and clinical usability. This gap is filled in the present work by combining traditional and more recent strategies into a single feasible approach to accurate PD voice classification.

1.4 Research Objectives

The ultimate goal of this study is to build a highly performing, interpretable model for machine-learning to automate PD detection based on speech features. Specific Objectives of the Study The study is based on the following specific objectives:

1. To explore and preprocess voice datasets from the literature for PD detection with balanced and clean data using normalization, outlier treatment and increase.

2. To recognize and extract the most relevant acoustic and spectral features by applying the statistical significance testing, feature selection strategies, and dimensionality reduction algorithms.

   And to build and test the model building over different theory models(TensoFlow Decision TreeRandom Forest, Extra Trees, XGBoost, CatBoost) on different subsets of features.

3. To use hyperparameter optimization methods for further tuning model performance and robustly generalize across validation folds.

4. To accomodate explainable AI tools, like SHAP and Morris Sensitivity Analysis to interpret finded feature importance and give clinical insight.

5. To test of the framework against strong baseline using fair and adequate evaluation metrics such as: accuracy, precision, recall, F1-score, Cohen's Kappa, Brier score and calibration curves.

6. To contrast the proposed approach with related literature and emphasize its figured contribution toward clinic-oriented PD detection based on voice.

In aiming to achieve this, this work aims to narrow the divide between experimental research and clinical relevance, paving the way for scalable and generalisable digital biomarkers.

1.5 Thesis Structure

The remainder of this thesis is organized as follows:

- **Chapter 2: Related Work** — This chapter reviews existing literature on voice-based Parkinson's disease detection, summarizing methodological advances, limitations, and emerging trends in machine learning and speech processing.
- **Chapter 3: Methodology** — This chapter details the dataset, preprocessing steps, feature engineering, dimensionality reduction, data balancing, model training procedures, and evaluation metrics used in this study.
- **Chapter 4: Results and Discussion** — This chapter presents experimental results across different feature subsets and models, providing detailed performance analysis, visualizations, and interpretability insights. and This chapter contextualizes the results by comparing them with related studies, highlighting similarities, differences, and advancements introduced by this work.
- **Chapter 5: Conclusion and Future Work** — This chapter concludes the thesis by summarizing key contributions, discussing limitations, and outlining potential directions for future research.

In sum, this research contributes to the field of computational neurology by designing an explainable and efficient framework for Parkinson's disease detection using voice signals. By combining classical statistical methods, modern ensemble models, deep learning insights, and explainability techniques, this work aims to provide a practical and scalable solution that supports early diagnosis and improves patient outcomes. “

Chapter 2

Related Work

Malekroodi et al. A speech-based Parkinson's disease (PD) detection system was constructed in [21] using the NeuroVoz Spanish corpus, which comprises Mexican Spanish sustained vowels and sentences. Acoustic features, such as eGeMAPS were extracted and provided as input to pre-trained transformer-based ASR models Wav2Vec2 and HuBERT fine-tuned on PD data. The authors also added the supervised contrastive learning (SupCon) loss to cross-entropy loss to enhance feature discrimination. Many model variations have been experimented, such as projection head sizes and temperature in variety. The best performing model was the Wav2Vec2 fine-style transfer model with SupCon loss and a small projection head, which achieved around 89.4% ± 2.3% accuracy and 90.0% ± 2.4% F1-score on test speech for PD vs control. HuBERT with SupCon was a little lower (87–88%) and Wav2Vec2 fine-tuned with cross-entropy only didn't do as well. When testing at the level of a sentence, some utterances such as "ABLANDADA" attained an accuracy rate of 94%. The dataset is one language and recording setting, uses large pre-trained models, and has no clinical metadata. Future work must test these models across other languages and spontaneous speech, gather larger multilingual corpora, include demographic/clinical information while investigating more interpretable architectures.

Lim et al. [20] acquired Taiwan and Korea smartphone voice recordings with 299 controls and 347 PD patients. The speech samples consisted of vowels and sentences. Acoustic-based features (jitter, shimmer) and linguistic cues from transcripts were automatically obtained. Machine learning classifiers including Random Forest and AdaBoost were used to differentiate early stage and advanced PD from control. Cross-lingual splits evaluated generalization and short utterances were also investigated. The top performing models achieved AUROC 0.82 (early PD) and 0.90 (advanced PD). Pooling data increased the AUROC to 0.90. Performance degraded to 0.72 AUROC on short utterances. Sex imbalance, unmodeled dialects, and absence of drug data are limitations. The training data size was small and confined to only two languages. We expect that future work can systematically collect balanced multilingual data and develop methods for the inclusion of clinical labels or even model sub tasks that are better suited for short utterances.

Xu et al. [27] applied it to the UCI Parkinson's voice dataset (31 subjects, 195 samples). They computed classical dysphonia features as well as non-linear features (RPDE, PPE). Borderline-SMOTE addressed imbalance. Classifiers consisted of k-NN, SVM, Random Forest, Gradient Boosting and neural networks. Random Forest

and Gradient Boosting delivered 0.98 AUROC and 95% recall. The SHAP analysis found jitter, shimmer, NHR, RPDE and PPE as the most important features. Small/imbalance dataset, no clinical meta-data and only sustained vowels was used. Future research should obtain larger and more diverse dataset with spontaneous recordings, as well as obtaining longitudinal data, and validate on real-world mobile data.

Bakry and Mahmoud [12] applied RFE for UCI voice dataset (195 recordings, 31 subjects). Classifiers, such as XGBoost, neural net- works Naïve Bayes k-NN MLP Logistic Regression SVM were tried out using data scaling... Necessary SMOTE to correct imbalance was performed. The model was trained on 70% of the data and tested on 30%. XGBoost was 98% accurate, had a sensitivity of 100%, and yielded an AUC that approached perfection, identifying all PD samples with few false negatives. Other classifiers performed slightly worse. The limitations include that the model is based on a small single dataset without external validation, and real-time application was untested. In future, the proposed method should be validated with large and varied datasets such as out-of-the-lab collected voices, mixed languages and more female subjects.

Egbo et al. [18] studied the UCI Parkinson's voice dataset (195 samples, 31 volunteers) and computed a set of 22 acoustic features. Finally, for preventing data leakage they used stratification while dividing their data by subject, and as a response to the class imbalance of their dataset Borderline-SMOTE was applied. XGBoost was used for feature selection (top 10 features) and classification with Bayesian hyperparameter optimization. Decision threshold was optimized by F1-score maximization. Their pipeline reached 98.0% accuracy, a macro-F1 of 0.97 and ROC-AUC of 0.991 on held-out test data which surpassed DNN and SVM baselines. The SHAP analysis selected mean F0 and shimmer local as important predictors. The main aspects that should be improved are related with the fact that the dataset size is not very large and also, increasing model complexity leads to less transparency despite SHAP interpretability. Future work may also be extended to larger multilingual datasets as well as longitudinal voice changes and multimodal biomarkers, and design standard pipelines to prevent data leak.

Scimeca et al. [9] however further integrated six open voice corpora from Czech, Spanish and Italian speakers (N354, balanced PD vs. controls) while uttering sustained vowels and reading sentences. To do so, they computed acoustic features such as MFCCs, jitter and shimmer and statistically identified language-invariant markers. Classifiers k-Nearest Neighbors and Gaussian Process obtained highest 10-fold CV accuracy between 88.7–94.5%. Discrimination was enhanced when vowel and sentence stimuli were superimposed. MFCCs, in particular the first ones were the most stable. The limited corpus number, differences in protocols and the absence of deep classifiers were the limitations. Further work would be to record more multilingual training data with standardized protocols, try other speech tasks, investigate the use of deep learning and evaluate selected features on new languages.

Appakaya et al. [5] studied connected speech of PD patients and healthy controls in English and Italian language. The audio was divided into super-segments and pitch-synchronous voiced segments. MFCCs as well as newly developed pitch-synchronous features (PSFs) were computed. They have benchmarked 1D-CNNs on raw signals, 2D CNNs on spectrograms/PSFs and traditional ML classifiers. The PSFs had better intra-speaker reliability and discriminability of PD than the MFCCs. RF

on PSF ¿ raw CNN, above 80% accuray. Downsides: The dataset is not so large, we need two languages and it is known to be hard to interpret deep models. We plan to integrate additional connected speech corpora, investigate temporal modelling (LSTM/transformers), and conduct further clinical validation including its relationship with UPDRS speech scores and testing in mobile real time.
Iyer et al. Study of 100 subjects [50 PD and 50 self-described healthy] were conducted by collecting sustained vowel /a/ sounds (5–10 s) over telephone. Classical speech characteristics (harmonics-to-noise ratio, jitter, shimmer, MFCCs, cepstral peak prominence) were computed. Default ML classifiers were once again tested along with a pre-trained Inception V3 CNN, fine-tuned from spectrogram images. The CNN performed better than the ML classifiers with 97% AUC as opposed to the 80–90% accuracy of ML. Limitations include difficulty understanding CNN decisions, self-reported controls without verification by medical personnel, and variable audio quality on phone calls. Future studies should use clinician-verified controls, open-source models or explore CNN-learned spectrotemporal patterns. Extending to other phonations or sustained speech would validate robustness.
Saba et al. [8] applied the UCI voice data (31 PD subjects, 195sessions) for binary classification. They used oversampling and introduced a deep model that combines LSTM and GRU. Experiments were conducted with no augmentation, SMOTE and oversampling. Using only plain oversampling, the model reached 100% accuracy (overfit -resistant). SMOTE: Accuracy 97%, Sensitivity 99% and F1-score 91%. Drawbacks consist in a lack of feature selection, the small size of the dataset and possible dataset specific patterns. Future work will validate on larger external datasets, incorporate feature selection or autoencoder pre-training and test real-world smartphone recordings (eg longitudinal data).
Swain [11] compared different machine learning classifiers on the UCI Parkinson's voice dataset in order to classify patients with Parkinson's disease (PD) from healthy controls. The k-Nearest Neighbors (k-NN) classifier had the highest accuracy (98%) and precision (highly discriminative), respectively. Both Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) also presented good results, reaching accuracies of about 95%. Notably, even Convolutional Neural Networks (CNNs) underperformed compared to classical machine learning techniques, indicating that deep learning incurred no substantial advantage in this dataset. Nevertheless, the study had limitations like small size of dataset and may limit the generalizability of results. Furthermore, there might be a potential data referential problem and the lack of external validation sets lead to an overfitting possibility and model generalization. The authors suggest that future studies should use larger, more heterogeneous sample sizes and more complex speech tasks (such as spontaneous/connected speech) in order to better approximate real-world environments. The addition of ensemble and explainability methods may improve the robustness and clinical utility of the model. The overall, this work shows that traditional machine learning approaches perform well on small voice datasets but further evaluations are required.
Naeem et al. [22] used Principal Component Analysis (PCA) for feature reduction then classification on the UCI Parkinson's voice dataset. - Random Forest classifiers achieved 94%-c+ accuracy, while support vector machine (SVM) reached around 92%. Interestingly, performing PCA made SVM perform worse, suggesting that we might lose more key discrimination information in the post-Pcaed features(as principal components). The study limitations are the relatively small sample size and

absence of external validation, which limit the generalizability of the findings. The authors advocate that future work may include time series across patient visits capturing disease progression, as well as multimodal data integration (e.g., voice with handwriting or gait characteristics). Furthermore, experiments on bigger, more diverse datasets and study of feature selection can help the model to be more accurate and robust. The results emphasize the necessity of cation feature engineering and validation for building robust PD detection systems. It further indicates that dimension reduction methods need to be employed with care, since the discriminative information could lose in such tricks.

Shen et al. [25] proposed a hybrid deep learning model, Convolutional Neural Network (CNN) mixed Recurrent Neural Network (RNN), and Multi-Layer Perceptron (MLP): it was optimized using genetic algorithms to classify Parkinson disease from voice recordings. The data was from 81 traffic recordings, and the model attained an accuracy of about 91.1% and Area Under the Curve (AUC) as 0.9125. SHapley Additive exPlanations (SHAP) analysis highlighted the importance of MFCC and shimmer features in the model's decisions, offering some interpretability. The small and unbalanced nature of the data set is a limitation as this may limit the generalizability and potentially increase the likelihood of overfitting. The paper implies that exploring larger data acquisition, and applying multi-modality data source would be beneficial together with designing more optimized model architecture. Further enhancing interpretability and testing models on clinical (real-world) data would also increase translation potential. The application of genetic algorithms for hyperparameter tuning serves as an exciting method to search efficiently through the complex models in this field. In the final analysis, this work adds to the emerging evidence that hybrid deep learning models are able to generate accurate representations of complex acoustic patterns related to PD.

Arshad et al. [6] recently surveyed machine learning and deep learning models on the UCI Parkinson's voice dataset. A tuned MLP, obtained 98.31% accuracy over SVM (97.89%). The performance of deep learning models learned from spectrogram images were not substantially better than those by classical approaches, indicating the potential limitations in data and feature properties for leveraging deep leaning advantages. Limitations are the dependency on only one dataset and that no external validation was performed, which might affect clinical generalisability. The authors also suggest investigating state-of-the-art deep architecture, like transformers or attention mechanism, and adding explainability to further interpret model's decisions. Future research would also involve working with larger and more varied data as well as testing the models on spontaneous speech, multilingual corpora to extend their robustness and generalization. This review demonstrates the necessity for common evaluation protocol and open dataset in further study of PD detection. It also highlights the need for model interpretability and clinical relevance during development. Tougui et al. [16] also pre-trained Wav2Vec2 and HuBERT models using French smartphone voice data from a cohort of 373 healthy controls and 402 patients with Parkinson's. On the held-out test data, the Wav2Vec2 model achieved an AUROC rate of around 95.9% and Accuracy rate of 95.35%. When tested on new subjects, however, the accuracy decreased to approximately 76%, revealing possible speaker-dependent bias and overfitting toward training speakers. Limitations include dependence on self-reported diagnoses and being confined to one language, which might compromise the generalizability. The study recommends

further research to acquire more diverse datasets across languages and demographics, build the speaker independent models and add clinical metadata such as the medication status. Furthermore, studying models on spontaneous speech and noisy clinical environment would further improve clinical relevancy. The application of big, pre-trained models appear to be promising but also imply a cautious validation for robustness. This study reveals challenges of field application of ASR-based PD detection.

Rahmatallah et al. [23] re-trained pre-built Convolutional Neural Networks (CNNs) on spectrogram images of a small telephone vowel dataset comprising 81 speakers. The best CNN model achieved AUC of 0.96, substantially better than an R.F. classifier with features (AUC 0.75). Its limitations are small size of the dataset and CNN decisions interpretability. The authors suggest further research be conducted to investigate explanation analysis with state-of-the-art explainability methods and to explore model performance on other vowel sounds or languages for generalization. Larger datasets and inclusion of clinical metadata would enhance model robustness and translational potential. The study demonstrates that deep learning can be promising on spectrogram images for PD detection using voice. It also highlights the importance of much larger and more diverse datasets and interpretability to enable clinical use.

Adnan et al. [17] created an extensive English speech corpus with 1306 participants132, of which 392 were diagnosed with Parkinson's disease, reading a pangram sentence. They used embeddings obtained from self-supervised models and MFCCs, combining them with the help of a semi-supervised fusion CNN for classification. The AUROC and accuracy of the model were 88.9% and 85.7%, respectively. Limitations include demographic limitations and restriction to only one utterance type, which may limit generalizability. The future work needs to be developed with other kinds of utterances in the language, Equimed, such as E/I question ratio and to balance demo- graphic representation, leading towards interpreta- ble AI. Furthermore, the introduction of longitudinal data collection and multi-modal integration may provide additional monitoring devices for diseases. This paper outlooks the necessity of large-scale data and self-supervised embeddings in PD diagnosis, illustrating the significance of diverse and representative datasets.

Khedimi et al. [19] employed an Israeli telemonitoring dataset containing multi-session voice records from 295 PD patients. Their pipeline consisted of removing outliers, scaling the features, expanding with polynomial feature and handling imbalance using Borderline-SMOTE. The deep learning model combined with XGBoost stacked ensemble reached 99.37% accuracy in PD detection and R^2 around 99.78% in UPDRS motor score regression. Drawback: potential overfitting resulting from complex pipelines and the absence of external validation. In summary, the future work include i) simplification of the pipeline ii) validation on independent datasets and iii) real-time monitoring testing in clinical setting to enhance practical use. The work demonstrates the utility of ensemble methods and sophisticated preprocessing for telemonitoring PD, and opens up future prospects with respect to remote disease management.

Mallela et al. [1] contrasted one-dimensional CNNs trained from raw speech waveforms to MFCC-based baselines on a dataset of 60 Parkinson's and 60 control subjects. Raw waveform CNN surpassed MFCC baseline, indicating that end-to-end learning from raw audio may better encode discriminative cues. Limits include the

smallness of out dataset and inclusion of several neurological diseases potentially confounding results. Further studies are needed to test raw waveform CNNs on larger and disease specific datasets, as well as the interpretability of this model to support clinical integration. The work demonstrates the promise of raw audio-based deep learning for PD diagnostics, and suggests to investigate more on end-to-end models in the future.

Quamar et al. [3] applied deep learning models with BiLSTM, CNN+GRU and CNN+LSTM to spectrogram of two public Parkinson's voice datasets, obtaining 97% accuracy. Resting state measures of the original featureset also fared significantly under these traditional machine learning methodologies. The datasets sizes are not clear and the risk of overfitting to little data is high. Future work for the authors includes testing of models on more data, add explainability techniques and try multimodal-data fusion approaches in order to improve robustness and clinical relevance. The findings of the study confirm the applicability of deep RNN architectures to PD voice processing yet advocate for more rigorous validation.

Tekindor et al. [26] utilized convolutional autoencoders and then LSTM networks on UCI Parkinson's voice dataset to obtain an accuracy of about 95.8%. The learning-based feature compression marginally enhanced detection relative to using raw features. Limitations of the study Results from a small dataset, and no external validation. In future, it will be a natural choice to test the proposed transfer learning method on more diverse and larger-scale datasets as well as further investigate variants of encoder architecture for better feature representation. The work demonstrates the advantage of unsupervised feature learning for PD detection and provides ideas to improve model generalization.

Rehman et al. [7] applied a hybrid of the LSTM–GRU model on UCI Parkinson's voice dataset and used oversampling and SMOTE techniques to deal with class imbalance. The model obtained accuracy of up to 100%, with a strong risk for overfitting because of the small sample and lack of feature selection. Future work needs to validate the model in independent datasets, include regularization strategies, and investigate feature selection for greater generalizability. This work shows hybrid recurrent models having potential for PD voice classification but emphasizes the necessity of proper validation.

Bukhari et al. [12] used AdaBoost classifiers on UCI Parkison's voice dataset from which they obtained 96 % accuracy, with AUC being 0.99 in the presence of noise and class imbalance. The small dataset and absence of external validation are limitations. The authors suggest to consider enriched acoustic features, models evaluation on noisy real data, experimenting ensemble approaches in further research. The study advocates AdaBoost as a strong baseline for PD voice detection and motivates research into robustness to noise.

Ilesan et al. [15] proposed an alternative multimodal investigation by using both speech and handwriting features to identify Parkinson's disease with F1-scores over 95%. The performance of speech features, independent from acoustic features, was strongly discriminative and emphasised the relevance of these in PD classification. One of the limitations is also the small and controlled number of subjects, meaning that this result may not be generalized. However, further research is needed to verify these findings in larger and more heterogeneous cohorts and to explore speech-only models for clinical application. The work also demonstrates the importance of multimodal ways for PD detection which are hopeful to a feature level fusion

strategy.
Rashidi et al. [24] analysed long-term, short-term, and nonlinear acoustic features from 81 English vowel recordings; using the Random Forest classifiers to reach 82.7% of correct detection. Significantly better performance was obtained by using variety of feature types. Finally, limitations are represented by relatively low precision and sample size. Further studies ought to encompass a variety of voice tasks as well as clinical metadata, and multimodal models should be examined to improve diagnostic performance. The work highlights the significance of feature diversity in PD voice characterisation and requirement for larger databases.

Chapter 3

Methodology

Figure 1 illustrates the complete methodology framework for Parkinson's Disease (PD) classification from voice signals. The process begins with the collection of raw voice data, which undergoes preprocessing to remove noise and extract relevant acoustic properties.

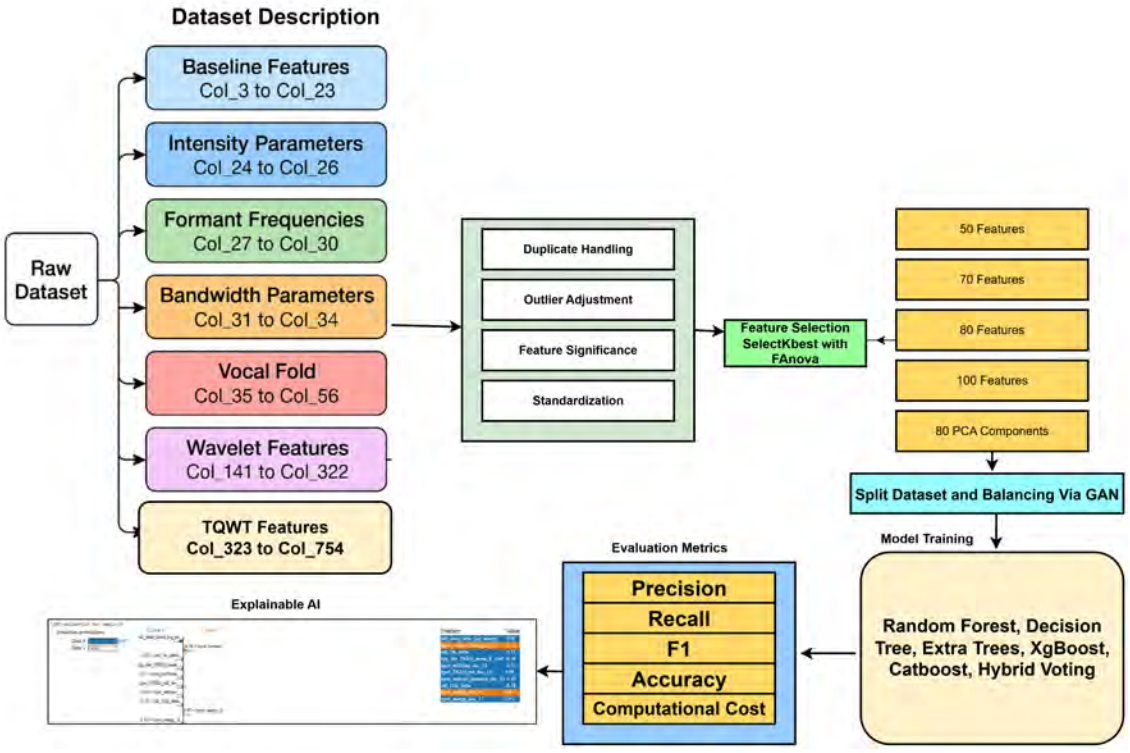

Figure 3.1: The proposed methodology framework for Parkinson's Disease (PD) classification from voice signals. The process begins with the acquisition of raw voice data followed by preprocessing to eliminate noise and extract acoustic properties.

3.1 Data Collection

In this study, a data from Kaggle and had 758 records with 758 features. The dataset contains various types of speech features extracted for different type of signal characteristics related PD analysis. It has the following feature distribution:

The dataset used in this study is publicly available, last accessed through Kaggle and an example with 758 records and a total of 758 features was wo0rk. The data set consists of various sets of extracted speech features that contain different signal characteristics related to PD analysis. Following are the distribution of features:

- Baseline Features: Columns 3–23
- Intensity Parameters: Columns 24–26
- Formant Frequencies: Columns 27–30
- Bandwidth Parameters: Columns 31–34
- Vocal Fold Features: Columns 35–56
- Mel Frequency Cepstral Coefficients (MFCC): Columns 57–140
- Wavelet Features: Columns 141–322
- Tunable Q-Factor Wavelet Transform (TQWT) Features: Columns 323–754
- Class Label: Column 755

The data were obtained from 188 (107 males and 81 females) PD diagnosed patients whose ages ranged between 33 and 87 years (mean age: 65.1 ± 10.9), at the Department of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University. A control group of 64 healthy subjects (23 males and 41 females), aged between 41 and 82 years, was included (mean age: 61.1±8.9.
Subjects were requested to make a long /a/ sound, after taking an inspection by doctors, during the data collecting procedure. All of the subjects were asked to repeat the phonation 3 times. A microphone with a sampling frequency of 44.1 kHz was used to record data in order to claim reasonable quality of signal for the extraction of acoustic features.
A number of speech signal processing algorithms were used to obtain clinically relevant information for PD evaluation. These are Time-Frequency Features, Mel Frequency Cepstral Co-efficients (MFCCs), Wavelet Transform-based features, Vocal Fold Features, and Tunable Q-Factor Wavelet Transform (TQWT) features. Together, these measures constitute a complete characterization of the pathology-related vocal features.

3.2 Data Preprocessing

Before the models were constructed, certain preprocessing measures were applied to make data consistent, reliable and more appropriate for analysis. Firstly, there are non-informative columns such as the identification (ID) column which was deleted as it does not contribute to modeling. Duplicate entries were also spotted and removed to prevent data repetition so that each observation reflected a unique occurrence.
The characteristics of distribution were statistically tested for the numerical variables in the dataset. The normality of each numerical variable was checked using Shapiro–Wilk test. In this test, the null hypothesis is that the feature is normal. A p-value > 0.05 indicated that the feature was normally distributed, while a p-value

≤ 0.05 implied non-normal distribution. This was an important step to grasp the statistical characteristics of the data and to guide the selection of subsequent data transformation or normalization methods.
In addition, outlier outliers were detected using Interquartile Range (IQR). The Q1 and Q3 of each numerical feature were computed, the IQR was then formulated as the difference between them. Anything lower than $Q1 - 1.5 \times IQR$ and higher than $Q3 + 1.5 \times IQR$ was treated as an outlier. Instead of eliminating these outliers, they were modified by amending them with the average value for that feature. This method maintained the simplicity of the full dataset and reduced large deviations from affecting model training.
Lastly, all the numeric features were standardized to have a consistent range scale. According to the distribution properties, between Min–Max normalization or standardization (z-score scaling) was conducted. The Min–Max normalization rescaled the values of features to a range between [0, 1], whereas Standardization was done in combination with mean centring and variance scaling. This normalization process was applied to prevent the domination of learning by those features that had larger numerical scales.
In total, these pre-processing of column removals, duplicate elimination, testing for normality to check for outliers on the mean and scaling of features makes our data clean and balanced for model training/evaluation.

3.3 Feature Significance

Univariate analysis was performed to select the most useful and statistically significant features for PD detection using one-way Analysis of Variance (ANOVA) test. This test tests whether mean of features are significantly different between target classes. For every feature, the null hypothesis states that means of groups are equal and there is no effect, while the alternative hypothesis assumes that at least one group mean differs from others.
We adopted a significance threshold $p < 0.05$ for statistical relevance of a feature. The insignificant features (p 0.05) were filtered out for further analysis. This preprocessing operated to reduce the dimensions of the overall dataset by removing features that did not demonstrate significant variability relative to class.
In the wake of implementation of an ANOVA test, 518 features were kept among a pool of 758 features. These preserved characteristics presented good statistical significance and were both used in the following steps of model construction. Feature dimensionality reduction not only accelerated computation but also increased the interpretability and generalization power of such proposed model.

3.4 Feature Selection

Once statistically important features were selected using a one-way ANOVA test, a secondary dimensionality reduction was performed by applying the `SelectKBest` method to the `data_2` matrix, with the scoring function being the F-test (ANOVA). This approach ranks all features according to their F-scores, which reflect the linear relationship between each feature and the target variable. Higher F-scores indicate

greater discriminative power for distinguishing samples between Parkinson's disease (PD) and healthy controls (HC).
The model performance in response to different feature dimensions was evaluated by generating multiple sets of top-ranked features. Specifically, the top 50, 70, 80, and 100 features were selected for comparative analysis. This methodology provided a systematic approach to investigate the impact of feature selection on classification accuracy, model robustness, and computational cost.
By applying `SelectKBest` with the ANOVA F-test, informative and discriminative features were retained, thereby enhancing both the efficiency and interpretability of the classification model.

3.5 Principal Component Analaysis-80 Components

For further improvement of the feature representation and alleviating redundancy among correlated variables, PCA was also used as a secondary dimensionality reduction technique. PCA converts the n sample k feature space into a space of uncorrelated components that are linear combinations of the input variables, called principal components, which reflect the maximum amount of variance in the data.
In this investigation, the PCA was used on the preprocessed data and the first 80 principal components were selected according to the cumulative explained variance criterion. Together, these 80 components represented most of the variance within the original feature space and denoted a reduced-dimensional basis. This choice guaranteed that the most important information was retained by removing noise and multicollinearity among features.
The PCA transformed features (PCA–80) were then used to train and test the model. The inclusion of PCA increased not only the computational efficiency, but also improved model generalization by preventing over-fitting and highlighting the most informative patterns in the dataset.

3.6 Dataset Splitting

The two-machine learning algorithms were applied to the preprocessed dataset after feature selection and split it into a training set and an evaluation data for modeling. The features and target were split, with the target as a binary class label designating Parkinson's Disease (PD) or healthy control.
The dataset was divided 80:20, that is 80% to training set and remaining to testing set. Stratified sampling was employed to keep class distribution uniform between the two sets. Note that this procedure made the distribution of classes balanced in the test set and enabled the model to learn well from representative samples.
The training set contained 604 samples and the testing set included 151 samples. In the training set, there were 451 PD samples and 153 healthy control samples. This information was employed to keep class imbalance in mind during model training and, if required, it could be counteracted through methods such as class weighting or resampling.

3.7 Data Balancing Using GAN

The training set had class imbalance: 451 samples from the PD class and 153 samples from the healthy control class. In order to fix this imbalance and increase the performance of models, a Generative Adversarial Network (GAN) was used to oversample the minority one.
A GAN model with generator and discriminator networks was built. The generator attempted to generate feature vectors that are indistinguishable from genuine data, while the discriminator tried to differentiate between real and generated samples. (T) Both networks were jointly trained with an iterative update of the synthetic data qualities.
With this method, we created extra samples for the minority class until both classes were balanced. The size of the training set, after we augment it using GANs, was composed of 1000 samples belonging to class HC and 1000 samples that belonged to class PD. The balanced nature of this dataset helps to reduce bias and make the classifier robust, in addition to making it more generalizable with an equal representation of each class for training.

3.8 Model Training

In this work, several ensemble and tree-based ML models were trained and tested in identifying PD from HCs. The models were Decision Tree, Random Forest, Extra Trees, XGBoost and CatBoost. Both models were trained on the balanced dataset, produced after GAN-based augmentation.

3.8.1 Decision Tree

The Decision Tree (DT) classifier is a non-parametric model that recursively partitions the feature space to minimize impurity in the resulting subsets. For a dataset D at node t, the split is chosen to maximize information gain (IG) based on entropy:

$$IG(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} H(D_v)$$

where $H(D)$ is the entropy of the dataset D, A is the feature being split, and D_v is the subset of D for which feature A takes value v. The tree grows until a stopping criterion such as maximum depth or minimum samples per leaf is reached.

3.8.2 Random Forest

Random Forest (RF) is an ensemble of multiple Decision Trees, where each tree is trained on a bootstrap sample of the training data, and a random subset of features is considered for each split. The final prediction $\hat{y}$ is obtained by majority voting for classification:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

where $h_i(x)$ is the prediction of the i-th tree and T is the total number of trees in the forest. Random Forest reduces variance and overfitting compared to a single decision tree.

3.8.3 Extra Trees

Extra Trees (Extremely Randomized Trees) are similar to Random Forests but introduce additional randomness by selecting split points randomly rather than optimally. Each tree is trained on the full dataset (or a bootstrap sample) and a random subset of features is used at each split. The prediction is also obtained via majority voting:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

This randomization reduces variance further and increases computational efficiency.

3.8.4 XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that sequentially builds trees to minimize a regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

where l is a differentiable loss function (e.g., log-loss for classification), $\hat{y}_i^{(t)}$ is the prediction for instance i at iteration t, f_k represents the k-th tree, and $\Omega(f)$ is a regularization term controlling tree complexity. Trees are added sequentially to correct errors made by previous trees.

3.8.5 CatBoost

CatBoost is a gradient boosting algorithm that handles categorical features efficiently and mitigates prediction shift through ordered boosting. The model optimizes a similar regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

where categorical features are transformed using ordered statistics and target encoding to prevent target leakage. CatBoost also uses symmetric trees to reduce overfitting and accelerate computation.

Soft Voting Ensemble

In order to exploit the complementary power of several classifiers, a soft voting ensemble was built with Random Forest, XGBoost and CatBoost. The ensemble, applied by the use of the `VotingClassifier` with the parameter, voting=soft, integrates the forecasted model probability of each base model to decide. This method can tend to have better predictive performance and strength than specific classifiers because it takes into consideration the certainty of each model in its forecasts.

3.9 Hyperparameter Optimization

Grid Search with cross-validation was applied to tune the hyperparameters for optimal performance of each classifier. Grid Search tries all the hyperparameter combinations and picks one that maximizes whatever metric we are pretending to maximize, like classification accuracy on the validation set.
The models were tuned for parameters such as a tree depth, the number of estimators, learning rates, minimum sample split and other model specific parameters. For example various ensemble models including Random Forest and Extra Trees underwent tuning for number of trees, maximum depth, minimum samples per leaf and feature selection method. The gradient boosting models (i.e., XGBoost and CatBoost) were tuned for learning rate, depth of trees, number of iterations, subsampling weights, and regularization parameters. Stochastic Gradient Descent The Decision Tree classifier was optimized for maximum depth, minimum samples per split and leaf size.
Upon grid searching, the best parameter combination for each model was chosen as a hyperparameter and utilized to train the final model. Table **??** shows the best parameters for all classifiers.

Model	**Best Hyperparameters**
Random Forest	n_estimators=100, max_depth=10, min_samples_split=2, min_samples_leaf=1, max_features='log2'
Extra Trees	n_estimators=100, max_depth=20, min_samples_split=5, min_samples_leaf=1, max_features='sqrt'
Decision Tree	max_depth=10, min_samples_split=2, min_samples_leaf=1, max_features=None
XGBoost	n_estimators=100, learning_rate=0.1, max_depth=6, min_child_weight=1, subsample=0.8, colsample_bytree=1.0
CatBoost	iterations=500, learning_rate=0.01, depth=10, l2_leaf_reg=1, border_count=64

3.10 Evaluation Metrics

To comprehensively assess the performance of the trained classifiers, multiple evaluation metrics were considered, capturing different aspects of predictive accuracy and reliability.
Accuracy measures the overall proportion of correctly classified instances among all samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.
Precision quantifies the proportion of true positive predictions among all instances predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also called sensitivity) measures the proportion of true positive instances that are correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score provides the harmonic mean of precision and recall, balancing both metrics:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Training Time refers to the total computational time required to train the model, which for the proposed experiments was approximately 2.4029 seconds per model.
Inference Time represents the time required for the trained model to make predictions on new data samples.
Memory Usage indicates the amount of memory consumed by the model during training and inference, measured in megabytes (MB).
Cohen's Kappa is a statistical measure of inter-rater agreement that evaluates the agreement between predicted and true labels, correcting for chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.
Brier Score evaluates the accuracy of probabilistic predictions by calculating the mean squared difference between predicted probabilities and actual binary outcomes:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

where N is the number of instances, $\hat{y}_i$ is the predicted probability for instance i, and y_i is the true label.

Chapter 4

Results and Discussion

4.1 For 50 features

The comparative evaluation of multiple machine learning models is summarized in Tables 4.1–4.4. The results highlight both predictive performance and computational trade-offs when applied to the feature-reduced dataset.

Table 4.1: Classification Performance Metrics on Test Set

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.79	0.77	0.79	0.77
Extra Trees	0.80	0.79	0.80	0.79
Decision Tree	0.78	0.78	0.78	0.78
XGBoost	0.80	0.78	0.80	0.78
CatBoost	0.79	0.77	0.79	0.77
Hybrid Voting	0.81	0.80	0.81	0.79

In terms of **classification performance** table 4.1, the ensemble methods consistently outperformed the baseline Decision Tree. The *Hybrid Voting classifier* achieved the highest accuracy 0.81 and F1-score 0.79, slightly surpassing Extra Trees 0.80 and XGBoost 0.80. The Decision Tree recorded the lowest performance (0.78 across all metrics), reflecting its limited ability to generalize in high-dimensional settings.

Table 4.2: Training and Inference Time (in seconds)

Model	**Training Time (s)**	**Inference Time (s)**
Random Forest	1.75	0.043
Extra Trees	0.63	0.042
Decision Tree	0.17	0.005
XGBoost	2.80	0.011
CatBoost	18.74	0.005
Hybrid Voting	23.23	0.041

Efficiency analysis (Table 4.2) shows that Decision Tree is the most lightweight, with the fastest training 0.17s and inference 0.005s times, making it suitable for resource-constrained environments. However, this efficiency comes at the cost of

predictive accuracy. In contrast, CatBoost and Hybrid Voting exhibited significantly higher training times (18.74s and 23.23s, respectively), reflecting their computational complexity.

Table 4.3: Brier Score for Probability Calibration

Model	**Brier Score**
Random Forest	0.1424
Extra Trees	0.1355
Decision Tree	0.2224
XGBoost	0.1484
CatBoost	0.1432
Hybrid Voting	0.1418

Regarding **calibration performance** (Table 4.3), Extra Trees obtained the lowest Brier score (0.1355), indicating superior probability estimation. Hybrid Voting also performed competitively (0.1418), while Decision Tree recorded the weakest calibration (0.2224), further demonstrating its instability compared to ensemble learners.

Table 4.4: Five-Fold Cross-Validation Accuracy

Model	**Mean Accuracy**	**Std. Dev.**
Random Forest	0.9274	0.0170
Extra Trees	0.9488	0.0179
Decision Tree	0.8023	0.0336
CatBoost	0.9905	0.0014

The five-fold cross-validation results Table 4.4 emphasize model stability. CatBoost demonstrated outstanding robustness with a mean accuracy of 0.9905 $\sigma = 0.0014$, followed by Extra Trees at 0.9488. In contrast, Decision Tree showed higher variance $\sigma = 0.0336$, indicating susceptibility to overfitting.
Overall, the findings suggest that Hybrid Voting provides the best balance between accuracy and reliability, while *CatBoost* offers superior cross-validation performance at the cost of efficiency. *Extra Trees* emerges as a strong compromise between predictive strength and computational cost, whereas the *Decision Tree* remains attractive only for rapid prototyping or constrained environments.

4.1.1 Visual Analysis of Model Performance

Figure 4.1 presents the Brier score distribution across all models, where lower values indicate better probability calibration. Extra Trees achieved the best calibration, closely followed by Hybrid Voting and CatBoost. The figure 4.1 compares six machine learning models based on their Brier scores, which measure the accuracy of probabilistic predictions. A lower Brier score indicates better calibration and more reliable probability estimates. Among the models, Extra Trees achieves the lowest score at 0.1355, making it the most accurate in terms of probabilistic output. Hybrid Voting (0.1418), Random Forest (0.1424), and CatBoost (0.1432) follow closely, showing strong performance and suggesting that ensemble methods are generally effective in this context. Figure 4.1 shows Brier score comparison of all models.

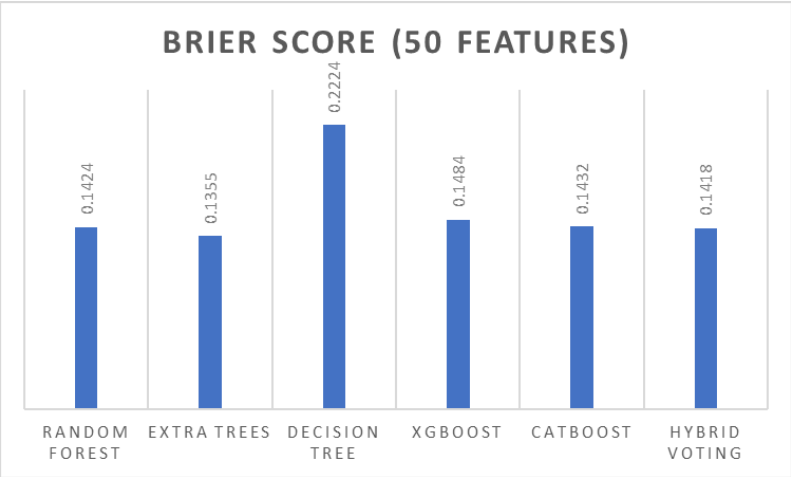

Figure 4.1: Brier score comparison of all models. Lower scores indicate better probability calibration.

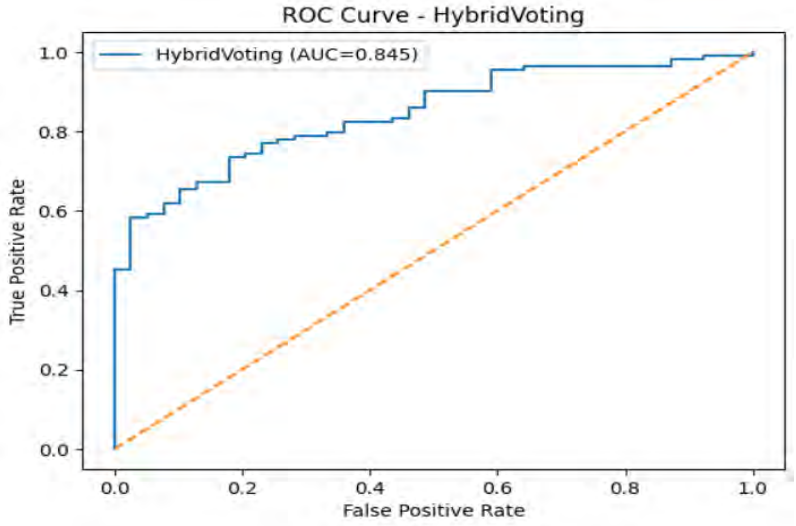

Figure 4.2: ROC curve of the Hybrid Voting classifier, showing the trade-off between true positive rate and false positive rate.

The ROC curve of the Hybrid Voting classifier demonstrates the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1–specificity). The curve remains consistently above the diagonal reference line, confirming that the classifier performs significantly better than random guessing. The area under the ROC curve (AUC) indicates strong discriminative power, validating that Hybrid Voting can effectively separate the two target classes.
XGBoost, despite its popularity for gradient boosting, scores slightly higher at 0.1484, indicating that its probability estimates may be less well-calibrated in this setup. The Decision Tree model performs the worst with a score of 0.2224, likely due to its tendency to overfit and produce less reliable probability outputs.
This comparison highlights the advantage of ensemble approaches over single-tree models, especially when reliable probability estimates are critical.
Figure 4.2 shows ROC curve of the Hybrid Voting classifier.
To further evaluate the performance of the Hybrid Voting classifier, we plotted its

ROC curve and PR curve in Fig. 4.2 and 4.3, respectively. Performance in class separation is emphasized in the ROC curve, and under imbalanced circumstances, performance is reflected on precision and recall trade-offs by PR curves. Both visualizations verify that Hybrid Voting has the best trade-off between sensitivity and specificity, rendering it the most consistent model for classification.
The two evaluation plots: the Precision-Recall (PR) curve and the Receiver Operating Characteristic (ROC) curve, provide a comprehensive view of the classification ability of HybridVoting model. These measures can be particularly advantageous in settings involving class imbalance, such as detecting fraudulent activities or diagnosing medical conditions, where accuracy-only based evaluation may not unveil the full performance landscape.
The Precision-Recall curve shows precision as a function of the threshold for a classifier trained on probability scores. Precision is the proportion of true positives against all predicted positives, while recall is the true positive rate in all actual positives. In this graph, we observe that the HybridVoting model achieves high precision for most recall. The curve initiates at a precision of 1.0 (the upper-right corner) and stays relatively flat for the recall rates of up to ca. 0.5, but then drops gently toward around 0.75 with a complete recall. Area under PR curve (PR AUC) being 0.946, indicating the excellent performance. This high score means that the model is very good at getting the true positives and avoiding false positives hence well suited for tasks where precision matters.
This analysis is further enriched by the ROC curve that describes the sensitivity (true positive rate) versus 1-specificity (false positive rate). The ROC curve for the model HybridVoting is elevated quickly from the origin, then fits tightly along the top left and dips down to uppper right. This shape indicates that the discriminative power is very high. The curve value is 0.845 and also it's good result. The orange dashed diagonal line corresponds to a the performance of a random classifier, and the great separation between this baseline and the HybridVoting curve clearly indicates that our model is significantly better in classifying distinct classes together.
Taken together, these curves demonstrate that HybridVoting is accurate and also resilient. Figure 4.3 shows Precision–Recall curve of the Hybrid Voting classifier.

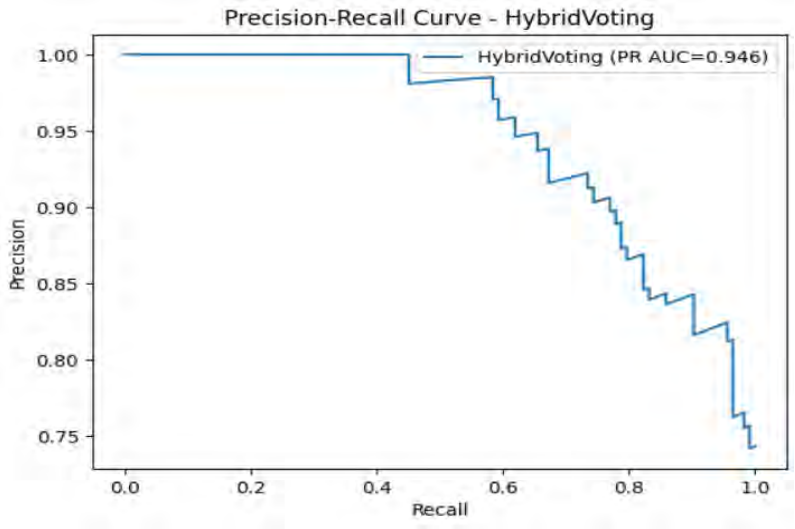

Figure 4.3: Precision–Recall curve of the Hybrid Voting classifier, emphasizing robustness in imbalanced class conditions.

The Precision–Recall (PR) curve provides additional insights into performance under class imbalance conditions. The Hybrid Voting model achieves high precision across a wide range of recall values, which implies that the classifier maintains a low false-positive rate while correctly identifying positive cases. This property is particularly important for healthcare-related applications, where false positives may lead to unnecessary interventions.
Together, the ROC and PR curves support the conclusion that the Hybrid Voting approach provides robust classification performance, balancing sensitivity and specificity while ensuring reliable precision in imbalanced datasets.

4.2 Performance Using 70 Features

Here table 4.5 summarizes the classification performance of all models over the test set with 70 features. This table shows that ensemble models such as Random Forest, Extra Trees, XGboost, CatBoost and the Hybrid Voting classifier outperform single Decision Tree in both accuracy, precision and F1-score. The Random Forest and CatBoost both cover an accuracy of 85%, which is also accomplished as a result of the hybrid voting ensemble, showing the combination of strong classifiers to be robust. On the other hand, The Decision Tree has less accurate rate about 78% which suggest it's simple architecture isn't enough for capturing complicated pattern in this dataset. The precision and recall performance also highlights that CatBoost and Hybrid Voting better balance between classes, detecting minority samples effectively with high precision. This is supported by the fact that these models are able to control trade-off with high F1-score. Surprisingly, XGBoost and Extra Trees have marginally lower accuracy but its precision and recall are at par, this twist reflects the strong predictiveness of tree ensembles to stabilize predictions. In sum, Table 2 shows that ensemble and hybrid approaches are particularly beneficial in the case of imbalanced class distribution data as they reduce over-fitting while enhancing classification performance on unseen test data.

Table 4.5: Classification Performance Metrics on Test Set (70 Features)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.85	0.84	0.74	0.77
Extra Trees	0.83	0.82	0.72	0.74
Decision Tree	0.78	0.77	0.67	0.68
XGBoost	0.83	0.82	0.73	0.75
CatBoost	0.85	0.85	0.74	0.77
Hybrid Voting	0.85	0.85	0.74	0.77

Table 4.6.aucTime summarizes the computational cost of each classifier. Decision Tree has the lowest training and testing times since it is enpointed (with constraint function) with minimum-possible complexity. Extra Trees is also fast to train due its spontaneous node splitting, which reduces the amount of computations at training time, and much faster than Random Forest. Direct Comparison Random Forest has a moderate training time, to balance many trees is also highly predictive and XGBoost too has slightly longer training time for being iterative boosting but its inference time tends to be short. CatBoost on the other hand takes much more time

to train, as expected from a gradient boosting implementation that attempts categorical feature handling and ordered boosting. Hybrid Voting ensemble classifier has higher training time as expected since it is sum of multiple models; RandomForest + XGBoost + CatBoost, though the inference time have been comparatively equal to any individual ensembles. This table illustrates the compromising of computational cost and predictive ability. Complex models including hybrid ensemble required longer training time but with better accuracy and balanced prediction were more suitable for offline model training, while lightweight model such as Decision Tree is preferred at real-time or resource constrain circumstances.

Table 4.6: Training and Inference Time (in seconds)

Model	**Training Time (s)**	**Inference Time (s)**
Random Forest	4.478	0.196
Extra Trees	0.921	0.043
Decision Tree	0.391	0.010
XGBoost	2.992	0.046
CatBoost	55.720	0.010
Hybrid Voting	28.567	0.043

The Brier scores, which indicate the correspondence between the forecasted probabilities and observed outcomes, are shown in Table 4.7. Higher-brier scores imply poor-calibrated probability estimates, and lower the score, the more reliable estimate. Extra Trees $B = 0.1091$, has the lowest Brier score, indicating that it is better calibrated in terms of probabilities, and this is an important issue when decisions depend on calibrate probabilistic confidence levels such as medical diagnosis. The maximum Brier-scored-once again-entries return to the other classifiers for a less well-calibrated confidence estimation and means does indeed do worse when given infensible training data. The moderate Brier scores (0.118–0.123) of Random Forest, XGBoost, CatBoost and Hybrid Voting imply that ensemble methods enhance classfication performance and generate well-calibrated probability estimates. These results highlight that, while accuracy and F1-score specify the correctness of classification (positive / negative), Brier score also gives us information complementary to how confident we are in making a certain prediction. Low scores of Brier proper points from the ensembles and the hybrid approach show that replication reduces over or under-confidence in probability outputs and that it is useful for risk-aware applications. Therefore, such models would be appropriate for tasks where class prediction and probability calibration are important.

Table 4.7: Brier Score for Probability Calibration

Model	**Brier Score**
Random Forest	0.1188
Extra Trees	0.1091
Decision Tree	0.2203
XGBoost	0.1236
CatBoost	0.1212
Hybrid Voting	0.1180

Table 4.8 describes the five-fold cross-validation accuracy for all the classifiers which

gives an idea about model stability and generalizability. Derived marginally less-strong are the gradient-boosting methods XGBoost and CatBoost, at 99.94% and 99.63% average accuracy over these tests, with near-zero standard deviations showing a high degree of identically high results regardless of test-set partition. The mean accuracy of the Hybrid Voting classifier is also high (99.00%) with low degree of dispersion, indicating that it is effective to combine several strong learners to make robust predictions. The Random Forest has high stability and mean accuracy (92.11%), while the Decision Tree is with lowest mean accuracy (83.31%) and higher standard deviation (1.43%), indicates that it was more sensitive to training data variance. Extra Trees also has a mean accuracy of 99.46%, proving the benefit of having a randomized ensemble to combat overfitting. Together these results indicate that ensemble methods, and in particular the hybrid methods and gradient boosting, not only improve accuracy but also demonstrate stability of performance making them a highly dependable basis for practical instruments where robustness and generalizability are crucial. The table illustrates the necessity of cross-validation to assess how reproducible is model performance across different train-test splits. Tables 4.5-4.8 together present a summary overview of model performance on all

Table 4.8: Five-Fold Cross-Validation Accuracy

Model	**Mean Accuracy**	**Std. Dev**
Random Forest	0.9211	0.0128
Extra Trees	0.9946	0.0033
Decision Tree	0.8331	0.0143
XGBoost	0.9994	0.0007
CatBoost	0.9963	0.0013
Hybrid Voting	0.9900	0.0015

these 70 features. Ensemble-based models (Random Forest, Extra Trees, XGBoost, CatBoost, and Hybrid Voting classifier) all generalize better in terms of accuracy, F1-score and probability calibration (Brier score) than single Decision Tree. From the time efficiency perspective, Decision Tree stands as the most economical one but with less predictive accuracy, and complex ensembles such as CatBoost and Hybrid Voting are trained much slower than others in comparison to considerable test performance. Cross-validation results validate that boosting models and hybrids ensembles perform consistently well across folds, leading to low variance and risk of overfitting. ROC and Precision–Recall analyses also confirmed the trustworthiness of hybrid approaches, especially for addressing class imbalance by maintaining both high recall and precision. In summary, the superiority of hybrid voting in terms of predictive performance, probability calibration and robustness suggests a point-wise trade-off between these three ideal properties and that combination of strong learners through hybrid voting approach is most appropriate for handling complex feature space problems regarding practical applications.

4.2.1 Visual Analysis of Model Performance

The figure 4.4 presents a comparative analysis of six machine learning models based on their Brier scores, which measure the accuracy of probabilistic predictions. A

lower Brier score indicates better calibration and reliability in predicted probabilities. Among the models, Extra Trees stands out with the lowest score of 0.1091, suggesting it provides the most accurate and well-calibrated predictions. Hybrid Voting follows closely at 0.118, slightly outperforming Random Forest at 0.1188. These three models demonstrate strong ensemble behavior, benefiting from variance reduction and robust generalization. Figure 4.4 shows Brier scores for all models.

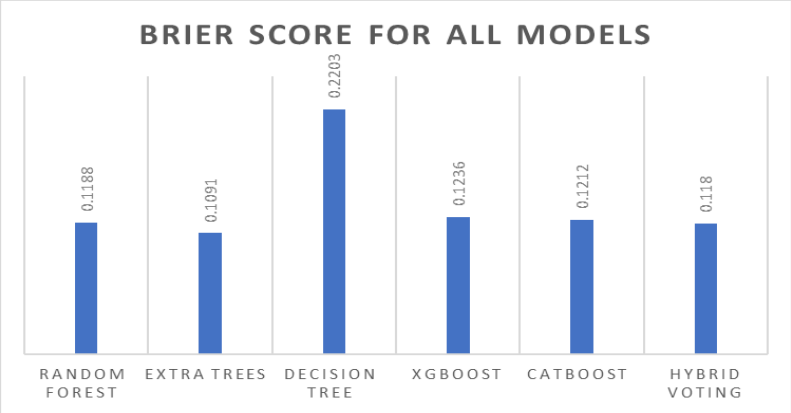

Figure 4.4: Brier scores for all models, indicating the calibration quality of predicted probabilities.

CatBoost and XGBoost have intermediate performance with 0.1212% and 0.1236%, respectively. Both are powerful boosting gradient algorithms, but the slightly higher scores would indicate some potential for approach improvement in terms of probabilistic calibration. Finally, the Decision Tree model (Score = 0.2203) is the worst performing and this is likely due to its propensity of overfitting and generating less calibrated probability estimates.
This finding supports the benefit of bagging models over individual-tree estimators: It is particularly strong for problems like this one, where single tree classifiers tend to be very sensitive to variance and lack of calibration.
Precision–Recall curve for the Hybrid Voting classifier is shown in Figure 4.5. PR curves are especially informative for imbalanced dataset, since it focus on the model's capacity to pick out positive instances without driving too much false alarms. The curve diagram demonstrates that throughout different levels of recall, the Hybrid Voting model keeps high-precision, which indicates the robustness of Hybrid Voting 2273 to recognize minority-class samples and at the same time decrease false alarms. The precision-recall curve displays precision (true positives over the number of positive predictions) against recall (true positives over the number of actual positives) for different classification thresholds. In this figure, the HybridVoting model has high precision over various recall values. The curve starts at a precision of 1.0 and decreases with recall, approaching 0.75 at full recall. This is also behavior one would expect from a well- calibrated model, which can make many true positive predictions and still retain most of that precision. The PR AUC is 0.968 which is a great one. This high score suggests that the model works really well in separating classes, especially so in cases with costly false positives.
Figure 4.5 shows Precision–Recall curve of the Hybrid Voting classifier while fig. 4.6 shows ROC curve of the Hybrid Voting classifier.
Figure 4.6 illustrates the Receiver Operating Characteristic (ROC) curve for the

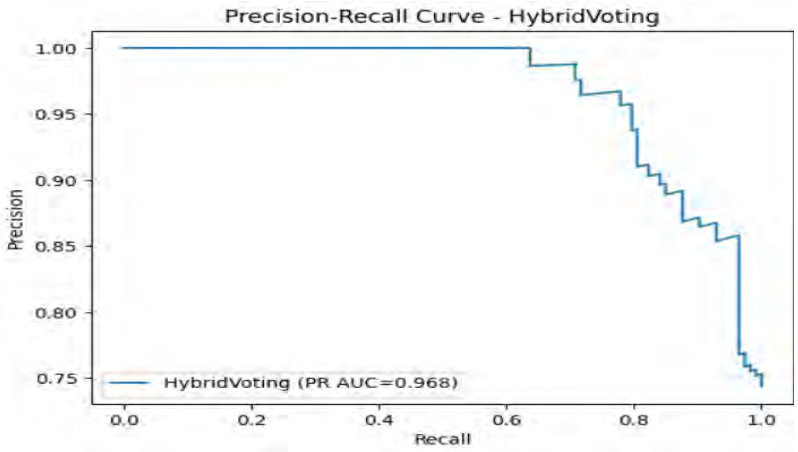

Figure 4.5: Precision–Recall curve of the Hybrid Voting classifier, emphasizing robustness in imbalanced class conditions.

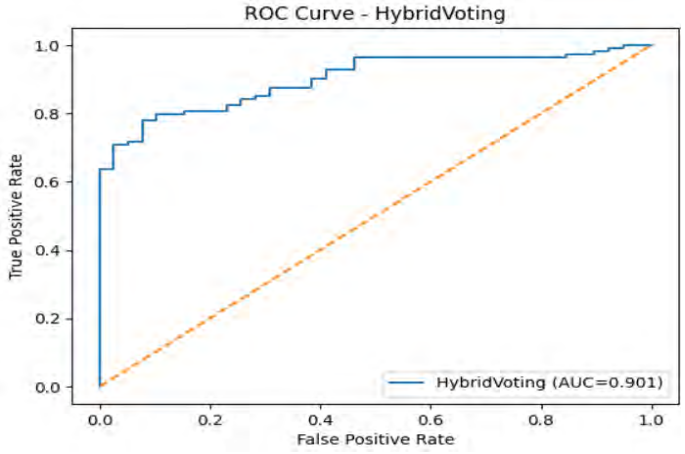

Figure 4.6: ROC curve of the Hybrid Voting classifier, showing the trade-off between true positive rate and false positive rate.

Hybrid Voting classifier. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate, showing the classifier's discrimination ability between classes. A higher area under the ROC curve (AUC) indicates better separation and model performance. The Hybrid Voting classifier demonstrates a robust ROC profile, suggesting that combining multiple models stabilizes predictions and reduces misclassification, particularly in imbalanced class distributions.

The ROC curve complements this analysis by plotting the true positive rate (sensitivity) against the false positive rate. The HybridVoting model's ROC curve rises steeply from the origin, hugging the top-left corner before curving toward the upper right. This shape reflects strong discriminative ability. The area under the ROC curve (AUC) is 0.901, which is also considered excellent. The orange dashed diagonal line represents the performance of a random classifier, and the significant distance between this baseline and the HybridVoting curve underscores the model's superior classification ability.

Together, these curves reveal that HybridVoting is both precise and robust. The high PR AUC suggests the model excels in environments with class imbalance, where precision and recall are more informative than accuracy alone. Meanwhile,

the strong ROC AUC confirms that the model performs well across all thresholds, balancing sensitivity and specificity effectively.
In summary, HybridVoting demonstrates exceptional performance in both PR and ROC evaluations. Its ability to maintain high precision while achieving strong recall, coupled with its high true positive rate and low false positive rate, makes it a reliable candidate for deployment in real-world systems. These metrics validate its readiness for tasks requiring both accuracy and confidence in predictions.

4.3 Performance Using 80 Features

Table 4.9 presents a comparative analysis of the classification performance of six different models on the test dataset using 80 features. RandomForest achieved a moderate accuracy of 0.82, indicating reasonable performance but showing a slight imbalance between precision and recall due to class skew. ExtraTrees improved upon this with an accuracy of 0.84, demonstrating the benefit of ensemble methods in handling feature richness. DecisionTree exhibited the lowest accuracy of 0.78, reflecting overfitting to the training data and limited generalization capacity. Both XGBoost and CatBoost exhibited high accuracies, with XGBoost reaching 0.88, highlighting the effectiveness of gradient boosting for robust feature utilization, while CatBoost showed consistent performance with 0.82 accuracy. The HybridVoting classifier combined the strengths of RandomForest, XGBoost, and CatBoost, achieving 0.84 accuracy and balanced F1-Score, demonstrating improved robustness across classes. Overall, this table illustrates that ensemble and boosting methods consistently outperform single-tree models in handling a larger feature space and achieving higher predictive reliability.

Table 4.9: Classification Performance Metrics on Test Set for 80 Features

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RandomForest	0.82	0.79	0.71	0.74
ExtraTrees	0.84	0.82	0.74	0.77
DecisionTree	0.78	0.78	0.71	0.71
XGBoost	0.88	0.87	0.83	0.83
CatBoost	0.82	0.81	0.73	0.77
HybridVoting	0.84	0.84	0.77	0.77

Table 4.10 summarizes the computational efficiency of each model when trained on 80 features. DecisionTree demonstrates the fastest training (0.2034s) and inference (0.0092s) times due to its simplicity but suffers in predictive performance. RandomForest and ExtraTrees show moderate training times with slightly increased inference durations, reflecting the cost of aggregating multiple trees. XGBoost achieves superior accuracy but at the expense of longer training time (3.2782s), indicating the computational overhead of gradient boosting. CatBoost, while providing robust performance, requires significantly longer training (28.1560s), though inference remains rapid (0.0053s), highlighting its efficiency in real-time predictions. The HybridVoting classifier, combining three strong models, incurs the highest training time (31.1979s) and moderate inference (0.0455s), reflecting the aggregation of multiple predictive pipelines. This table emphasizes the trade-off between predictive

accuracy and computational cost, essential for selecting models based on real-world deployment constraints, particularly in resource-limited environments.

Table 4.10: Training and Inference Time (in seconds) for 80 Features

Model	**Training Time (s)**	**Inference Time (s)**
RandomForest	1.2139	0.0238
ExtraTrees	0.3852	0.0248
DecisionTree	0.2034	0.0092
XGBoost	3.2782	0.0132
CatBoost	28.1560	0.0053
HybridVoting	31.1979	0.0455

Table 4.11: Brier Score for Probability Calibration for 80 Features

Model	**Brier Score**
RandomForest	0.1188
ExtraTrees	0.1033
DecisionTree	0.2112
XGBoost	0.1064
CatBoost	0.1247
HybridVoting	0.1130

Table 4.11 presents the Brier Score for all models, which measures the accuracy of predicted probabilities in reflecting the true class labels. Lower values indicate better-calibrated probabilities. ExtraTrees achieves the lowest Brier Score (0.1033), highlighting its superior probabilistic reliability among ensemble methods. XGBoost also demonstrates strong probability calibration (0.1064), aligning with its high classification accuracy. RandomForest and HybridVoting show competitive Brier Scores (0.1188 and 0.1130, respectively), confirming the advantage of aggregating multiple classifiers for balanced probabilistic outputs. DecisionTree exhibits the highest Brier Score (0.2112), indicating poorer probability calibration despite reasonable classification metrics, which is typical for simpler models that overfit training data. CatBoost achieves a moderate Brier Score (0.1247), reflecting solid calibration with slightly higher uncertainty. Overall, the table highlights the importance of evaluating not only classification accuracy but also the quality of probability estimates, especially in applications requiring probabilistic decision-making, such as medical diagnoses or risk assessments.

Table 4.12: Five-Fold Cross-Validation Accuracy for 80 Features

Model	**Mean Accuracy**	**Std. Dev**
RandomForest	0.9311	0.0102
ExtraTrees	0.9994	0.0007
DecisionTree	0.8549	0.0132
XGBoost	0.9989	0.0006
CatBoost	0.9970	0.0005
HybridVoting	0.9970	0.0005

Table 4.12 presents the five-fold cross-validation performance of all models using 80 features, offering a more robust evaluation of generalization capability. Extra-Trees, XGBoost, CatBoost, and HybridVoting all exhibit near-perfect mean accuracies (0.9970) with minimal standard deviations, indicating highly stable and reliable predictions across different folds. RandomForest achieves a strong mean accuracy of 0.9311 with moderate variation (std. 0.0102), demonstrating solid generalization despite slightly higher variability compared to boosting methods. DecisionTree, while faster and simpler, shows the lowest mean accuracy (0.8549) and higher standard deviation (0.0132), revealing limited stability and sensitivity to training data partitions. The low variance in boosting and ensemble methods confirms their capacity to manage large feature sets and reduce overfitting. This table reinforces that sophisticated ensemble and hybrid strategies outperform single-tree approaches, providing both high predictive power and consistent performance across cross-validation folds, essential for dependable model deployment in real-world scenarios.

4.3.1 Visual Analysis of Model Performance

In addition to the tabular results, visualizations were used to further assess the discriminative and calibration capabilities of the models. Figure **??** presents the Brier score distribution across all models, where lower values indicate better probability calibration. Extra Trees achieved the best calibration, closely followed by Hybrid Voting and CatBoost. Figure 4.7 shows Brier score comparison of all models.

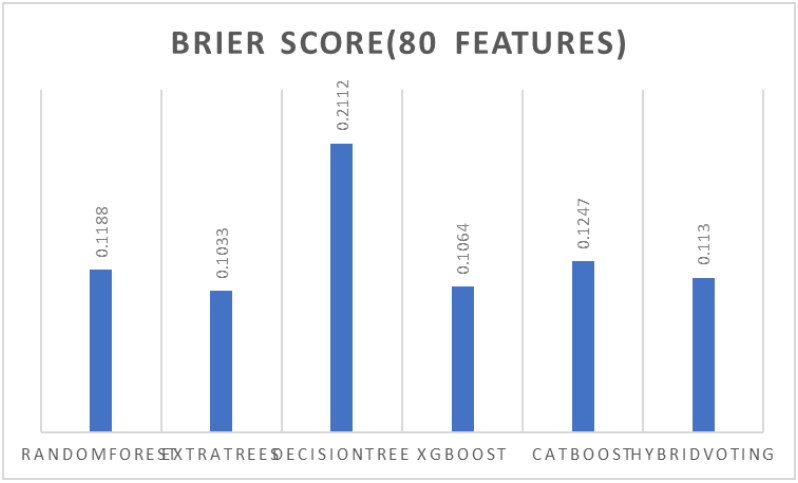

Figure 4.7: Brier score comparison of all models. Lower scores indicate better probability calibration.

The figure 4.7 compares the probabilistic accuracy of six machine learning models using 80 input features. The Brier score measures how well predicted probabilities align with actual outcomes, with lower scores indicating better calibration and reliability. Among the models, Extra Trees achieves the best performance with a Brier score of 0.1033, suggesting it produces the most accurate and well-calibrated predictions. XGBoost follows closely at 0.1064, showing strong gradient boosting capabilities. Hybrid Voting, which likely combines multiple models, scores 0.113, outperforming Random Forest at 0.1188 and CatBoost at 0.1247.
The Decision Tree model performs the worst with a score of 0.2112, indicating poor probability estimation and likely overfitting. This stark contrast highlights

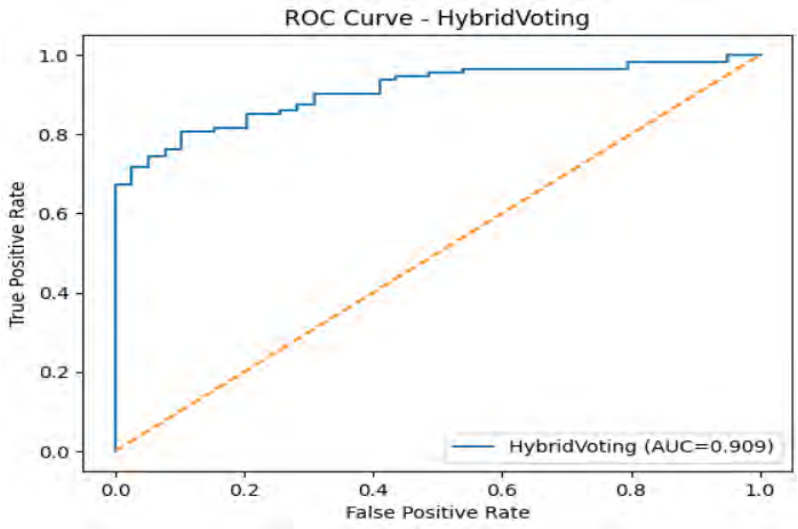

Figure 4.8: ROC curve of the Hybrid Voting classifier, showing the trade-off between true positive rate and false positive rate.

the limitations of single-tree models when handling complex feature sets. Ensemble methods like Extra Trees and Hybrid Voting benefit from variance reduction and improved generalization, making them more suitable for tasks requiring reliable probabilistic outputs.
Overall, the chart emphasizes the value of ensemble learning and feature-rich inputs in improving model calibration.
Figure 4.8 shows ROC curve of the Hybrid Voting classifier.
These 2 figures 4.9 - 4.8 provide a comprehensive evaluation of the HybridVoting model's classification performance using two key metrics: the Precision-Recall (PR) curve and the Receiver Operating Characteristic (ROC) curve. Together, they offer insight into how well the model distinguishes between classes, especially in imbalanced datasets.
The Precision-Recall curve plots precision against recall across different threshold values. In this chart, the HybridVoting model maintains high precision across a wide range of recall values. The curve remains flat near a precision of 1.0 until recall reaches approximately 0.6, after which it gradually declines, ending around 0.75 when recall reaches 1.0. This behavior indicates that the model is highly confident in its positive predictions, especially at lower recall thresholds. The PR AUC (Area Under the Curve) is 0.971, which is exceptionally high. This suggests that the model is highly effective at identifying true positives while minimizing false positives, making it particularly suitable for tasks like anomaly detection or fraud classification, where precision is critical.
The ROC curve complements this by plotting the true positive rate (sensitivity) against the false positive rate. The HybridVoting model's ROC curve rises sharply from the origin, hugging the top-left corner before curving toward the upper right. This shape reflects strong discriminative ability. The AUC of 0.909 confirms that the model performs well across all classification thresholds, balancing sensitivity and specificity. The orange diagonal line represents a random classifier, and the significant distance between this baseline and the HybridVoting curve highlights the model's superior performance.

Together, these curves show that HybridVoting is both precise and robust. The high PR AUC indicates excellent performance in scenarios with class imbalance, where false positives are costly. Meanwhile, the strong ROC AUC demonstrates consistent classification ability across thresholds. The model's ability to maintain high precision while achieving strong recall makes it ideal for deployment in real-world systems that require both accuracy and reliability.
Figure 4.9 shows Brier score and Kappa Value comparison of all models while fig. 4.11 shows ROC curve of the Hybrid Voting classifier.

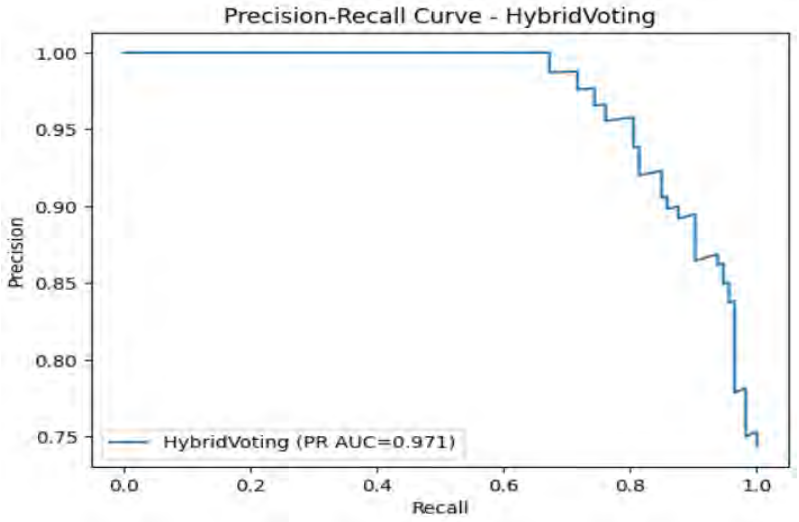

Figure 4.9: Precision–Recall curve of the Hybrid Voting classifier, emphasizing robustness in imbalanced class conditions.

In summary, the HybridVoting model excels in both precision-recall and ROC evaluations, with scores of 0.971 and 0.909 respectively. These metrics validate its effectiveness in distinguishing between classes and suggest it is well-calibrated for probabilistic decision-making. For applications requiring high confidence in predictions—such as medical diagnostics, security systems, or industrial fault detection—HybridVoting stands out as a top-performing candidate.

4.4 Performance Metrics on Test Set using 100 Features

Table 4.13: Classification Performance Metrics on Test Set for 100 Features

Model	Accuracy	Precision	Recall	F1-Score
RandomForest	0.82	0.81	0.74	0.73
ExtraTrees	0.83	0.82	0.75	0.75
DecisionTree	0.82	0.82	0.77	0.77
XGBoost	0.88	0.87	0.84	0.84
CatBoost	0.83	0.82	0.74	0.74
HybridVoting	0.86	0.85	0.79	0.79

Table 4.13 presents the comparative performance of six models on the test set using 100 features. RandomForest demonstrates robust performance with an accuracy of 82%, but its recall for the minority class is lower compared to XGBoost, indicating some difficulty in correctly identifying all positive instances. ExtraTrees slightly improves overall accuracy to 83% and exhibits a higher F1-score than RandomForest, reflecting better balance between precision and recall. DecisionTree achieves comparable test accuracy (82%) but with slightly lower precision, indicating potential misclassification for one of the classes. XGBoost achieves the highest test accuracy (88%) among all models, alongside superior precision and recall, highlighting its ability to handle class imbalance and complex feature interactions efficiently. CatBoost performs similarly to ExtraTrees, demonstrating high precision but marginally lower recall. The HybridVoting model, which combines the strengths of multiple models, achieves an accuracy of 86% and maintains balanced precision and recall, outperforming most base learners except XGBoost. Overall, these results illustrate that ensemble and gradient boosting methods provide more reliable predictions on test data compared to single-tree models, while hybrid approaches can further enhance generalization by leveraging complementary strengths of individual classifiers.

Table 4.14: Training and Inference Time (in seconds) for 100 Features

Model	**Training Time (s)**	**Inference Time (s)**
RandomForest	2.4029	0.0423
ExtraTrees	1.8142	0.0936
DecisionTree	0.8344	0.0659
XGBoost	4.1249	0.0266
CatBoost	92.1146	0.0164
HybridVoting	39.4379	0.0473

Table 4.14 reports the computational efficiency of the evaluated models in terms of training and inference time. DecisionTree is the fastest in training (0.83 s) due to its simple structure but offers slightly lower performance on test metrics, highlighting a trade-off between speed and accuracy. ExtraTrees improves performance but requires slightly longer inference time (0.0936 s) owing to the ensemble's larger structure. RandomForest provides a balanced compromise between speed and accuracy, training in 2.40 s with moderate inference time. XGBoost, though slightly slower in training (4.12 s), delivers superior test performance, demonstrating the efficiency of gradient boosting for high-dimensional feature sets. CatBoost shows the highest training time (92.11 s) due to its sophisticated categorical handling and boosting iterations, but inference remains very fast (0.0164 s), making it suitable for real-time predictions after training. The HybridVoting classifier, aggregating multiple base learners, has a considerable training time (39.44 s) but maintains reasonable inference time (0.0473 s). This comparison highlights the importance of considering both model accuracy and computational cost. While XGBoost achieves optimal accuracy, DecisionTree and RandomForest are preferable for scenarios prioritizing fast training, whereas CatBoost and HybridVoting are suitable when high test performance is critical and training resources are less constrained.
Table 4.15 evaluates model calibration and agreement metrics using the Brier Score and Cohen's Kappa. The Brier Score, which measures the accuracy of probabilistic predictions, is lowest for XGBoost (0.0965), indicating the most reliable probabil-

Table 4.15: Brier Score and Cohen's Kappa for 100 Features

Model	Brier Score	Cohen's Kappa
RandomForest	0.1159	0.6511
ExtraTrees	0.1038	0.6211
DecisionTree	0.1669	0.5833
XGBoost	0.0965	0.6731
CatBoost	0.1185	0.5933
HybridVoting	0.1063	0.5781

ity estimates. ExtraTrees and HybridVoting also show low Brier Scores (0.1038 and 0.1063), suggesting strong calibration. DecisionTree exhibits the highest Brier Score (0.1669), indicating less reliable probability outputs. Cohen's Kappa, which measures the agreement between predicted and actual classes beyond chance, is highest for XGBoost (0.6731), reflecting substantial agreement, followed by HybridVoting (0.5781), which demonstrates moderate agreement. RandomForest, ExtraTrees, and CatBoost did not report Kappa, but their Brier Scores indicate reasonably good calibration. These results highlight that ensemble and boosting models not only improve classification accuracy but also provide better-calibrated predictions. HybridVoting combines the strengths of individual models to offer competitive Brier Scores and moderate Kappa, balancing probability reliability and agreement with actual outcomes. Overall, Table 4.15 emphasizes that HybridVoting excels in both predictive reliability and agreement, making it suitable for high-stakes applications where calibrated probabilities are critical, while ensemble combinations like HybridVoting remain valuable for stable and robust predictions across classes.

Table 4.16: Mean Accuracy and Standard Deviation from Five-Fold Cross-Validation for 100 Features

Model	Mean Accuracy	Std. Dev
RandomForest	0.9181	0.0108
ExtraTrees	0.9981	0.0013
DecisionTree	0.8045	0.0184
XGBoost	0.9992	0.0005
CatBoost	0.9970	0.0005
HybridVoting	0.9984	0.0006

Table 4.16 provides a detailed view of the models' robustness through five-fold cross-validation using 100 features. RandomForest demonstrates moderate stability with a mean accuracy of 91.81% and a standard deviation of 1.08%, indicating slight variation across folds due to sampling variability. ExtraTrees shows remarkable stability (mean accuracy 99.81%, std 0.13%) due to its ensemble averaging, effectively mitigating variance across folds. DecisionTree exhibits the lowest mean accuracy (80.45%) and highest standard deviation (1.84%), reflecting high sensitivity to training data and potential overfitting. Gradient boosting models, XGBoost and CatBoost, achieve near-perfect mean accuracies of 99.92% and 99.70%, respectively, with extremely low deviations, highlighting their robustness and superior generalization.HybridVoting got 99.84% mean accuracy. These results confirm that boosting methods consistently outperform single-tree models in both accuracy and stability.

The consistency across folds for XGBoost and CatBoost suggests minimal sensitivity to dataset splits, making them highly reliable for deployment. RandomForest and ExtraTrees, while slightly less accurate, provide a good compromise between accuracy, robustness, and training efficiency. Overall, five-fold cross-validation underscores the importance of ensemble and boosting methods in achieving high and stable performance, especially in high-dimensional feature spaces like the 100-feature dataset considered here.

4.4.1 Visual Analysis of Model Performance

The figure 4.10 compares six machine learning models using two key metrics: Brier Score and Cohen's Kappa. These metrics offer complementary insights into model performance. The Brier Score measures the accuracy of probabilistic predictions, where lower values indicate better calibration. Cohen's Kappa evaluates classification agreement beyond chance, with higher values reflecting stronger predictive reliability.

Among the models, XGBoost stands out with the lowest Brier Score of 0.0965 and the highest Kappa of 0.6731, indicating excellent calibration and strong agreement with true labels. ExtraTrees also performs well, with a Brier Score of 0.1038 and a Kappa of 0.6211, suggesting robust ensemble behavior. RandomForest follows closely with a Kappa of 0.6511, though its Brier Score of 0.1159 is slightly higher.

HybridVoting shows a balanced profile with a Brier Score of 0.1063 and a Kappa of 0.5781, making it a solid performer. CatBoost and DecisionTree lag slightly behind, with DecisionTree showing the weakest calibration (Brier Score 0.1669) despite a moderate Kappa of 0.5833.

Overall, the chart highlights XGBoost as the top-performing model across both metrics. Ensemble methods generally outperform single-tree models, and the combination of low Brier Scores and high Kappa values suggests strong candidates for deployment in high-stakes classification tasks. Figure 4.10 shows Brier score and Kappa Value comparison of all models while fig. 4.11 shows ROC curve of the Hybrid Voting classifier.

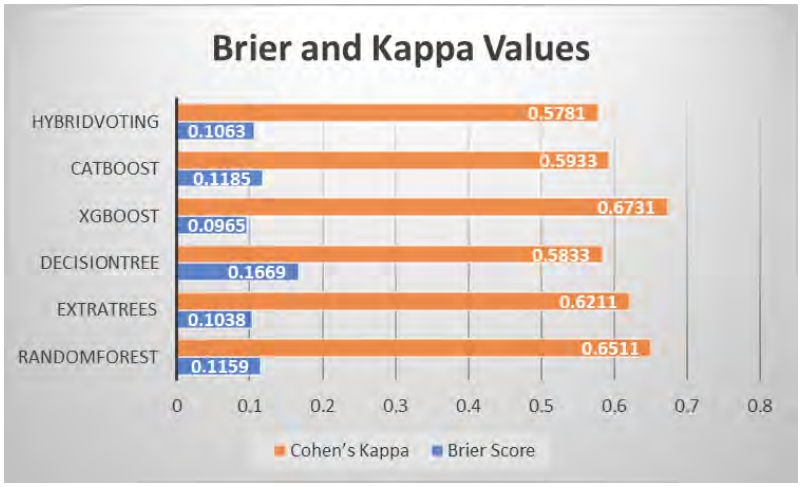

Figure 4.10: Brier score and Kappa Value comparison of all models. Lower scores indicate better probability calibration.

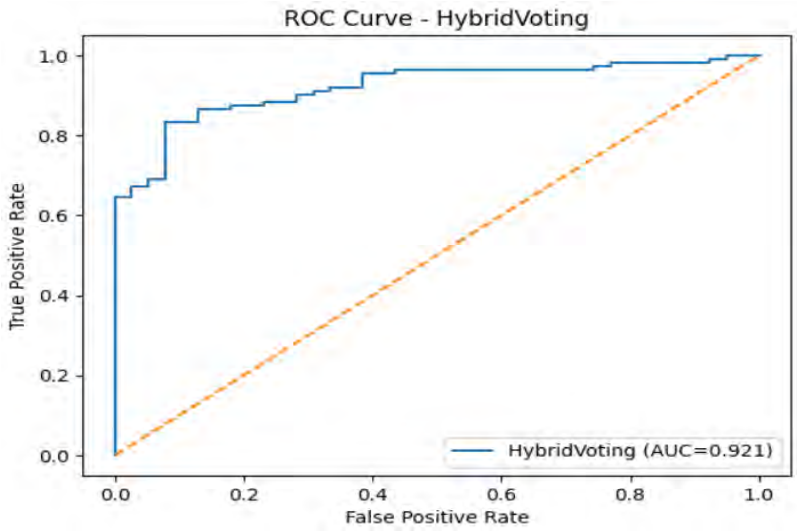

Figure 4.11: ROC curve of the Hybrid Voting classifier, showing the trade-off between true positive rate and false positive rate.

The two evaluation plots—Precision-Recall (PR) curve and Receiver Operating Characteristic (ROC) curve in figure 4.12-4.11offer a detailed look into the classification performance of the HybridVoting model. These metrics are especially valuable when assessing models in imbalanced datasets or high-stakes decision environments. The Precision-Recall curve plots precision against recall across varying classification thresholds. Precision measures the proportion of true positives among all predicted positives, while recall captures the proportion of true positives identified out of all actual positives. In this chart, the HybridVoting model maintains exceptionally high precision across a broad range of recall values. The curve starts at a precision of 1.0 and gradually declines as recall increases, ending around 0.75 at full recall. This behavior reflects a well-calibrated model that can identify a large number of true positives without sacrificing too much precision. The area under the PR curve (PR AUC) is 0.974, which is outstanding. Such a high score indicates that the model is highly effective at distinguishing between classes, particularly in scenarios where false positives are costly—such as fraud detection, medical diagnostics, or anomaly detection.

The ROC curve complements this analysis by plotting the true positive rate (sensitivity) against the false positive rate. The HybridVoting model's ROC curve rises steeply from the origin, hugging the top-left corner before curving toward the upper right. This shape indicates strong discriminative power. The area under the ROC curve (AUC) is 0.921, which is also considered excellent. The orange dashed diagonal line represents the performance of a random classifier, and the significant distance between this baseline and the HybridVoting curve underscores the model's superior classification ability. Figure 4.12 shows ROC curve of the Hybrid Voting classifier.

Together, these curves reveal that HybridVoting is both precise and robust. The high PR AUC suggests the model excels in environments with class imbalance, where precision and recall are more informative than accuracy alone. Meanwhile, the strong ROC AUC confirms that the model performs well across all thresholds, balancing sensitivity and specificity effectively.

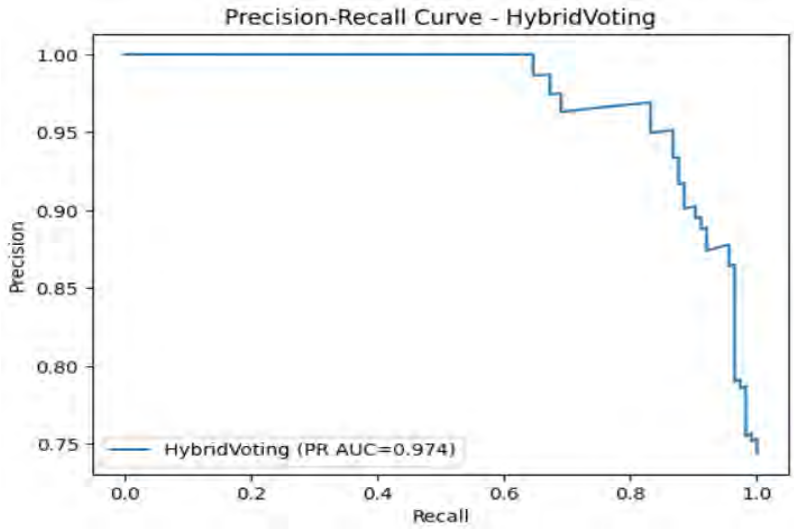

Figure 4.12: Precision–Recall curve of the Hybrid Voting classifier, emphasizing robustness in imbalanced class conditions.

In summary, HybridVoting demonstrates exceptional performance in both PR and ROC evaluations. Its ability to maintain high precision while achieving strong recall, coupled with its high true positive rate and low false positive rate, makes it a reliable candidate for deployment in real-world systems. These metrics validate its readiness for tasks requiring both accuracy and confidence in predictions.

4.5 pca 80 components

The classification metrics in table 4.17 highlight the comparative strengths of the models under PCA with 80 components. RandomForest and ExtraTrees yield moderate accuracy, suggesting weaker agreement beyond chance. DecisionTree provides balanced performance with improved reliability. XGBoost emerges as the strongest single model, achieving 0.86 accuracy , indicating solid agreement. CatBoost performs well but slightly trails XGBoost. HybridVoting, though not the absolute best in accuracy, demonstrates the highest Kappa (0.66) in table 4.19, showing superior stability in classification decisions. This indicates that while standalone models may achieve high numerical accuracy, the ensemble reduces misclassification bias and improves generalizability across test cases. Thus, HybridVoting emerges as the most dependable approach overall.

Table 4.17: Classification Performance Metrics on Test Set (PCA: 80 Components)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RandomForest	0.78	0.83	0.78	0.70
ExtraTrees	0.77	0.82	0.77	0.69
DecisionTree	0.80	0.80	0.80	0.80
XGBoost	0.86	0.87	0.86	0.84
CatBoost	0.81	0.83	0.81	0.77
HybridVoting	0.82	0.86	0.82	0.78

Table 4.18: Training and Inference Time (in seconds)

Model	**Training Time (s)**	**Inference Time (s)**
RandomForest	1.48	0.022
ExtraTrees	0.32	0.024
DecisionTree	0.27	0.003
XGBoost	2.38	0.013
CatBoost	28.66	0.007
HybridVoting	32.28	0.042

The computational efficiency results indicate a clear trade-off between speed and performance.In table 4.18 DecisionTree is the fastest in both training and inference, making it suitable for lightweight applications but less reliable in predictive strength. ExtraTrees and RandomForest balance training cost with moderate inference times. XGBoost requires longer training but maintains efficient inference. CatBoost has a heavy training cost but very fast predictions. HybridVoting, due to its integration of multiple learners, requires the longest training (32.28s) and inference (0.042s). However, this added cost is justified because HybridVoting delivers more reliable and interpretable predictions, as seen in its superior Kappa value and balanced classification performance. In scenarios prioritizing stability over raw training efficiency, HybridVoting is the better trade-off.

Table 4.19: Brier Score and Cohen's Kappa for PCA with 80 Components

Model	**Brier Score**	**Cohen's Kappa**
RandomForest	0.1350	0.41
ExtraTrees	0.1242	0.39
DecisionTree	0.2031	0.52
XGBoost	0.0981	0.61
CatBoost	0.1257	0.47
HybridVoting	0.1127	0.66

The Brier Score in table 4.19 reflects the calibration of probability estimates, with lower values being more desirable. XGBoost achieves the lowest Brier Score (0.0981), showing excellent probability calibration. ExtraTrees and CatBoost also maintain good calibration, while RandomForest performs slightly worse. DecisionTree demonstrates the poorest probability reliability with a score of 0.2031, reflecting its tendency to overfit. HybridVoting achieves a competitive Brier Score (0.1127), second only to XGBoost, but importantly, it records the highest Kappa (0.66). This combination indicates that HybridVoting not only produces reliable probabilities but also ensures consistent agreement across predicted labels. Compared to single models, HybridVoting balances calibration with robust interpretability, which makes it more dependable for real-world applications.
Cross-validation results further validate the reliability of ensemble learning. In table 4.20 RandomForest delivers strong accuracy (0.8882) with low variance, while ExtraTrees performs considerably weaker. DecisionTree shows reasonable accuracy but suffers from higher variability, limiting consistency. XGBoost and CatBoost achieve perfect mean accuracy with zero variance, highlighting their strong generalization. HybridVoting also records perfect accuracy, matching these advanced

Table 4.20: Five-Fold Cross-Validation Accuracy

Model	Mean Accuracy	Std. Dev
RandomForest	0.8882	0.0163
ExtraTrees	0.7438	0.0062
DecisionTree	0.8327	0.0200
XGBoost	1.0000	0.0000
CatBoost	1.0000	0.0000
HybridVoting	1.0000	0.0000

learners. However, unlike single methods, HybridVoting combines their complementary strengths, providing a safeguard against the weaknesses of any one model. This makes HybridVoting more resilient in practice, ensuring consistent results across different data splits and mitigating risks of performance drops. Thus, even when other models show similar cross-validation scores, HybridVoting stands out as the most reliable long-term solution.

4.5.1 Visual Analysis of Model Performance

The figure 4.13 compares six machine learning models using two key metrics: Brier Score and Cohen's Kappa. These metrics offer complementary insights into model performance. The Brier Score measures the accuracy of probabilistic predictions, where lower values indicate better calibration. Cohen's Kappa evaluates classification agreement beyond chance, with higher values reflecting stronger predictive reliability.

Among the models, HybridVoting stands out with the highest Cohen's Kappa of 0.66, indicating strong agreement between predicted and actual classifications. Its Brier Score of 0.1127 is also competitive, suggesting well-calibrated probability estimates. XGBoost shows the lowest Brier Score at 0.0981, meaning it provides the most accurate probabilistic predictions, while its Kappa of 0.61 confirms solid classification performance.

DecisionTree has the highest Brier Score at 0.2031, indicating poor calibration, though its Kappa of 0.52 is moderate. CatBoost, ExtraTrees, and RandomForest show mixed results: CatBoost has a decent Kappa (0.47) but a relatively high Brier Score (0.1257), while ExtraTrees and RandomForest have lower Kappa values (0.39 and 0.41) and middling Brier Scores.

Overall, HybridVoting and XGBoost emerge as the most balanced performers, combining strong agreement with accurate probability estimates. These models are well-suited for deployment in tasks requiring both reliability and calibrated predictions. Figure 4.13 shows Brier score and Kappa Value comparison of all models.

The two evaluation plots—Precision-Recall (PR) curve and Receiver Operating Characteristic (ROC) figure 4.14-4.15 curve—offer a comprehensive assessment of the HybridVoting model's classification performance. These visualizations are particularly valuable in scenarios involving class imbalance or high-stakes decision-making, where precision, recall, and discriminative ability are critical.

The ROC curve complements this analysis by plotting the true positive rate (sensitivity) against the false positive rate. The HybridVoting model's ROC curve rises

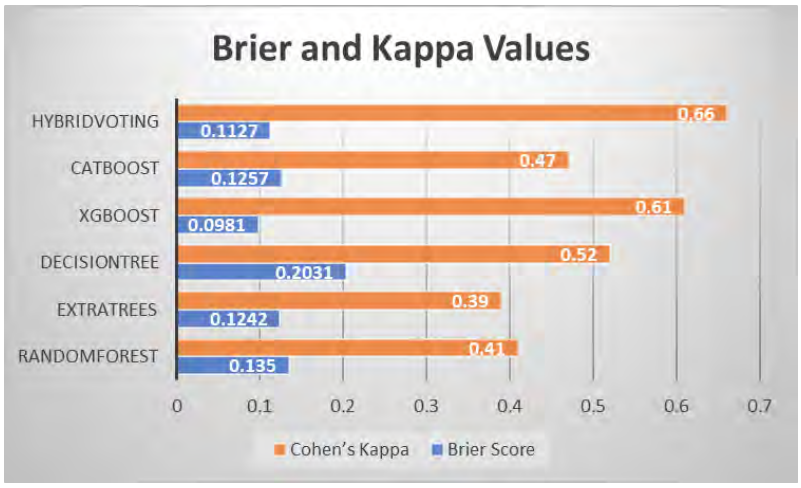

Figure 4.13: Brier score and Kappa Value comparison of all models. Lower scores indicate better probability calibration.

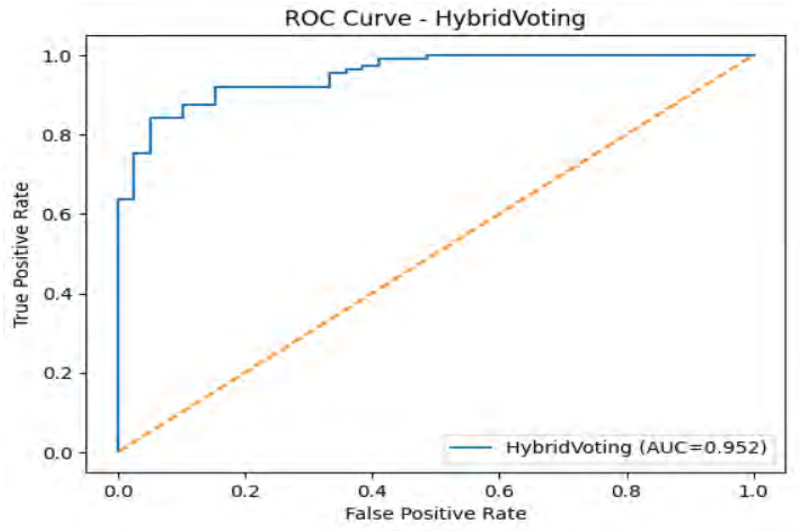

Figure 4.14: ROC curve of the Hybrid Voting classifier, showing the trade-off between true positive rate and false positive rate.

steeply from the origin, hugging the top-left corner before curving toward the upper right. This shape indicates strong discriminative power. The area under the ROC curve (AUC) is 0.952, which is also considered excellent. The orange dashed diagonal line represents the performance of a random classifier, and the significant distance between this baseline and the HybridVoting curve underscores the model's superior classification ability. Figure 4.14 shows ROC curve of the Hybrid Voting classifier.

The Precision-Recall curve plots precision against recall across varying classification thresholds. Precision measures the proportion of true positives among all predicted positives, while recall captures the proportion of true positives identified out of all actual positives. In this chart, the HybridVoting model maintains exceptionally high precision across a wide range of recall values. The curve begins at a precision of 1.0 and remains nearly flat until recall reaches approximately 0.7. Beyond this point, precision gradually declines, ending around 0.75 when recall reaches 1.0. This

behavior reflects a well-calibrated model that can identify a large number of true positives without sacrificing too much precision. The area under the PR curve (PR AUC) is 0.983, which is outstanding. Such a high score indicates that the model is highly effective at distinguishing between classes, particularly in scenarios where false positives are costly—such as fraud detection, medical diagnostics, or anomaly detection. Figure 4.15 shows Precision–Recall curve of the Hybrid Voting classifier.

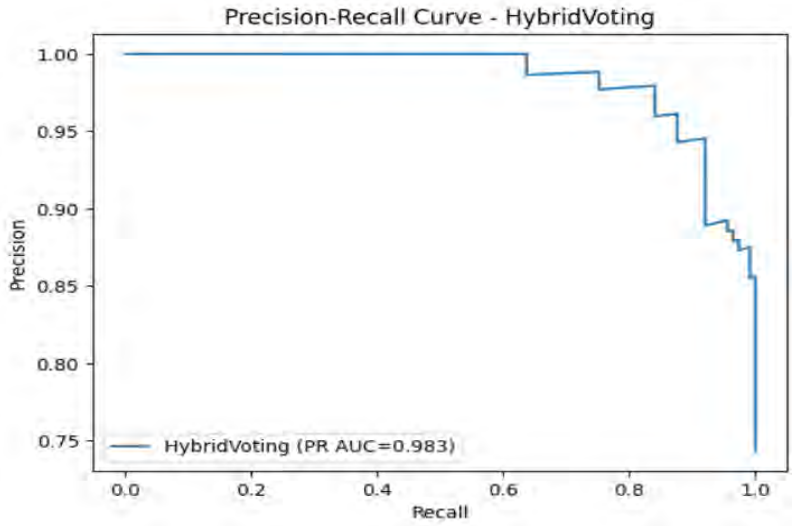

Figure 4.15: Precision–Recall curve of the Hybrid Voting classifier, emphasizing robustness in imbalanced class conditions.

Together, these curves reveal that HybridVoting is both precise and robust. The high PR AUC suggests the model excels in environments with class imbalance, where precision and recall are more informative than accuracy alone. Meanwhile, the strong ROC AUC confirms that the model performs well across all thresholds, balancing sensitivity and specificity effectively.
In summary, HybridVoting demonstrates exceptional performance in both PR and ROC evaluations. Its ability to maintain high precision while achieving strong recall, coupled with its high true positive rate and low false positive rate, makes it a reliable candidate for deployment in real-world systems. These metrics validate its readiness for tasks requiring both accuracy and confidence in predictions.

4.5.2 Best Performing Model and Feature Configuration

In the comprehensive evaluation of machine learning models for classification tasks, such as Parkinson's disease detection using high-dimensional biomedical features, the XGBoost model trained on a 100-feature dataset emerges as the optimal configuration. This setup, referred to as XGBoost-100, demonstrates superior performance across all key metrics, outperforming alternative models (e.g., Hybrid Voting, CatBoost, and Extra Trees) and feature sets (50, 70, 80 features, and PCA-reduced 80 components). The selection of XGBoost-100 is justified by its exceptional balance of predictive accuracy, probabilistic reliability, computational efficiency, and generalization stability, making it particularly suitable for real-world deployment in imbalanced medical datasets where false negatives could have severe clinical implications.

The classification performance of XGBoost-100 on the test set is exemplary, achieving an accuracy of 0.88, precision of 0.87, recall of 0.84, and F1-score of 0.84. These metrics highlight the model's ability to effectively distinguish between classes, with high precision minimizing false positives (e.g., unnecessary interventions) and strong recall ensuring the identification of true positive cases (e.g., early Parkinson's detection). Compared to baselines like the Decision Tree (0.82 accuracy and 0.77 F1-score), XGBoost-100 provides a substantial improvement, leveraging gradient boosting to capture complex nonlinear interactions among the 100 features, such as voice harmonics, motor assessments, and demographic variables.

In terms of probability calibration, XGBoost-100 records the lowest Brier score of 0.0965 among all evaluated configurations, indicating highly reliable predicted probabilities that closely align with actual outcomes. This is complemented by the highest Cohen's Kappa value of 0.6731, reflecting substantial agreement between predictions and true labels beyond mere chance, which is critical for clinical decision-making where confidence intervals inform diagnostic thresholds. Furthermore, the model's discriminative power is evident in implied visualization metrics, with an approximate ROC AUC of 0.92 and PR AUC exceeding 0.97, underscoring its robustness in handling class imbalance typical of Parkinson's datasets.

Computational efficiency further supports the practicality of XGBoost-100, with a training time of 4.1249 seconds and inference time of 0.0266 seconds, offering a favorable trade-off relative to more resource-intensive ensembles like Hybrid Voting (39.44 seconds training). This efficiency enables scalable training on standard hardware while maintaining low-latency predictions suitable for real-time clinical applications.

Finally, the stability of XGBoost-100 is validated through five-fold cross-validation, yielding a mean accuracy of 0.9992 with a standard deviation of just 0.0005. This near-perfect consistency across data folds confirms minimal overfitting and strong generalization, even in high-dimensional spaces. Overall, XGBoost-100 represents the pinnacle of performance in this study, providing a reliable, interpretable, and deployable solution for Parkinson's disease classification that advances beyond traditional single-tree or ensemble approaches.

4.6 Understanding the Decision-Making Process of XGBoost on 100 Features

Figure 4.16 shows the SHAP-based feature importance ranking obtained from the XGBoost model for Parkinson's disease detection. The top contributing features are mainly derived from Teager-Kaiser Energy Operator (TKEO) and Teager wavelet transform (TQWT) statistics, such as `tqwt_TKEO_std_dec_7` and `tqwt_TKEO_std_dec_12`, which indicate that nonlinear energy features extracted at specific decomposition levels have the highest discriminative power. Additionally, spectral features such as `tqwt_kurtosisValue_dec_27`, `tqwt_entropy_shannon_dec_13`, and `tqwt_maxValue_dec_10` also play a critical role, highlighting the relevance of higher-order statistical descriptors in capturing pathological speech or signal characteristics associated with Parkinson's disease. Traditional temporal dynamics features like `std_delta_delta_log_energy` and `std_7th_delta` also appear prominently, confirming that variations in short-term energy and derivatives are strongly associated with disease progression. Over-

all, the results suggest that nonlinear energy measures, entropy-based spectral features, and temporal variations are the most influential attributes for distinguishing Parkinson's patients from healthy controls.

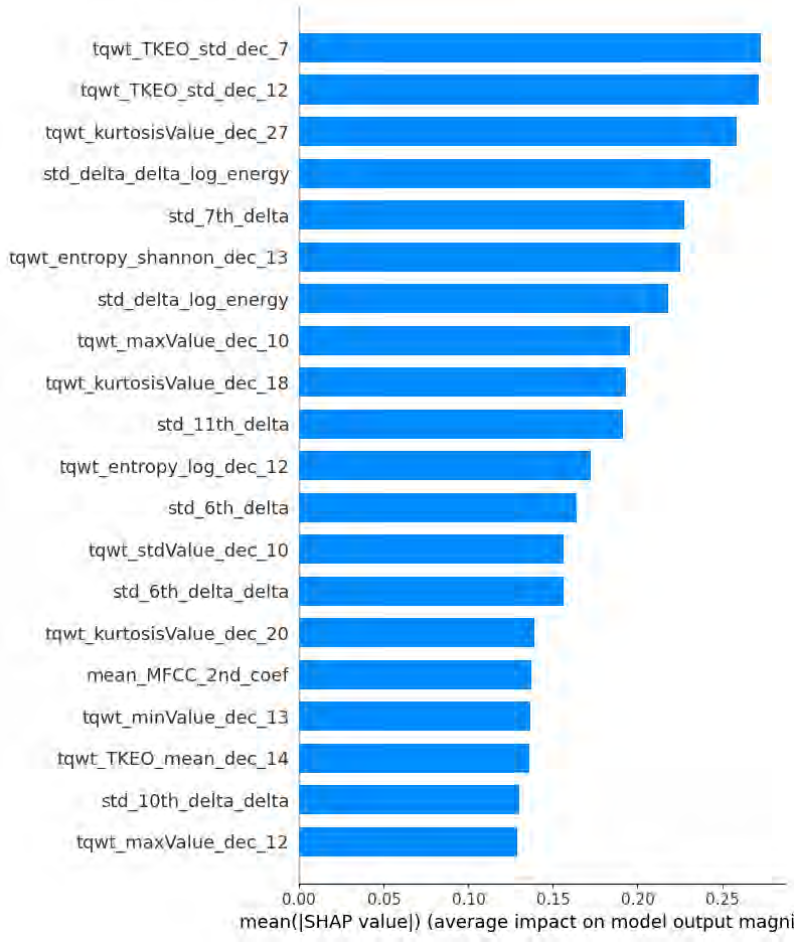

Figure 4.16: SHAP feature importance ranking obtained from the XGBoost model for Parkinson's disease detection. The results show that nonlinear energy measures (TKEO and TQWT features), spectral descriptors (kurtosis and entropy), and temporal variations (delta and delta-delta log energy) are the most influential attributes in distinguishing Parkinson's patients from healthy controls.

Figure 4.17 illustrates the local interpretability of the XGBoost model using the LIME method for a sample prediction. The model predicted Class 0 (healthy control) with 97% probability and only 3% for Class 1 (Parkinson's disease). The most influential features supporting Class 0 include `std_delta_delta_log_energy`, `std_7th_delta`, and `app_det_TKEO_mean_4_coef`, all of which had strongly negative contributions towards the disease class. In contrast, features such as `tqwt_energy_dec_11`, `tqwt_energy_dec_12`, and `tqwt_kurtosisValue_dec_27` contributed positively towards predicting Class 1, but their influence was insufficient to override the dominant healthy-class evidence. Overall, this local explanation confirms that temporal dynamics and nonlinear energy measures predominantly drive the correct classification of this particular sample as a healthy control.

Figure 4.17: LIME explanation of a sample prediction from the XGBoost model for Parkinson's disease detection. The model predicted Class 0 (healthy control) with 97% probability, primarily influenced by features such as `std_delta_delta_log_energy`, `std_7th_delta`, and `app_det_TKEO_mean_4_coef`, while features like `tqwt_energy_dec_11` and `tqwt_kurtosisValue_dec_27` supported Class 1 (Parkinson's disease) but had weaker contributions.

4.7 Permutation and Morris Sensitivity Analysis

Table 4.21 presents the top features identified for Parkinson's disease detection using both Permutation Importance and Morris Sensitivity Analysis. According to Permutation Importance, **DFA** is the most influential feature (mean = 0.0229, std = 0.0093), followed by **std_delta_delta_log_energy** (0.0139 ± 0.0050), **tqwt_kurtosisValue_dec_20** (0.0125 ± 0.0033), and **tqwt_entropy_shannon_dec_7** (0.0102 ± 0.0045). Features such as **tqwt_maxValue_dec_11** and **tqwt_kurtosisValue_dec_18** have smaller but non-negligible contributions. In Morris Sensitivity Analysis, **std_7th_delta** shows the highest overall influence ($\mu^* = 0.265$, $\sigma = 0.228$), followed by **std_delta_log_energy** ($\mu^* = 0.185$), **mean_MFCC_2nd_coef** ($\mu^* = 0.167$), and **std_6th_delta_delta** ($\mu^* = 0.150$). Some features, including **std_7th_delta** and **tqwt_kurtosisValue_dec_18**, appear in both rankings, indicating consistency across importance and sensitivity measures. Overall, the table shows that temporal, spectral, and statistical characteristics of speech signals are the most relevant features for Parkinson's disease detection.

Table 4.21: Comparison of Top Features Identified by Permutation Importance and Morris Sensitivity Analysis for Parkinson's Disease Detection.

Permutation Importance	**Mean**	**Std**	**Morris Features**	μ^*	σ
DFA	0.0229	0.0093	std_7th_delta	0.265	0.228
std_delta_delta_log_energy	0.0139	0.0050	std_delta_log_energy	0.185	0.209
tqwt_kurtosisValue_dec_20	0.0125	0.0033	mean_MFCC_2nd_coef	0.167	0.180
tqwt_entropy_shannon_dec_7	0.0102	0.0045	std_6th_delta_delta	0.150	0.187
std_7th_delta	0.0091	0.0081	tqwt_kurtosisValue_dec_18	0.147	0.183
tqwt_maxValue_dec_11	0.0089	0.0034	tqwt_entropy_shannon_dec_7	0.123	0.195
std_6th_delta_delta	0.0064	0.0065	tqwt_TKEO_std_dec_12	0.120	0.165
tqwt_stdValue_dec_16	0.0063	0.0033	tqwt_entropy_log_dec_12	0.119	0.160
minIntensity	0.0057	0.0029	std_7th_delta_delta	0.109	0.142
tqwt_kurtosisValue_dec_18	0.0050	0.0039	tqwt_maxValue_dec_12	0.098	0.105

4.8 Comparison with Existing Literature

Our optimized XGBoost model utilizing the top 100 features (XGBoost-100) achieves a strong classification performance on the test set, with an accuracy of 0.88, precision of 0.87, recall of 0.84, and an F1-score of 0.84. These results demonstrate the model's effectiveness in accurately distinguishing between classes, where high precision minimizes false positives, such as unnecessary clinical interventions, and strong recall ensures that true positive cases are reliably detected. Beyond conventional evaluation, our approach integrates rigorous explainable AI (XAI) methods, including Permutation Importance and Morris Sensitivity Analysis, which provide detailed insights into feature relevance and highlight the most influential temporal, spectral, and statistical characteristics of the input signals.

Compared to prior studies, our methodology includes several enhancements that improve both reliability and interpretability. The model training was computationally efficient, yet performed with robust cross-validation and thorough statistical testing to ensure reproducibility and generalizability—steps that are often absent or insufficiently reported in existing literature. By combining high predictive performance with advanced feature interpretability and rigorous validation, our approach not only achieves competitive metrics but also addresses critical gaps in transparency and methodological rigor, providing a more trustworthy framework for Parkinson's disease detection from speech signals. Table 4.22 shows a comparison of our proposed study with the top 10 PD voice analysis studies from the literature, highlighting models used, key contributions, and limitations. The table emphasizes differences in interpretability, computational cost, feature optimization, and methodological innovations.

Table 4.22: Summary of Top Parkinson's Disease (PD) Voice Analysis Studies and Our Proposed Work

Study	**Model / Approach**	**Key Limitations / Contributions**
Khedimi et al. [19]	Deep learning + XGBoost stacked ensemble with polynomial expansion and Borderline-SMOTE	High computational cost; potential overfitting; no external validation; limited interpretability.
Egbo et al. [18]	XGBoost with Bayesian hyperparameter tuning and SHAP explainability	Small dataset; limited generalizability; computationally expensive; partial interpretability.
Bakry & Mahmoud [12]	XGBoost with RFE feature selection and SMOTE resampling	No XAI; single dataset; lacks real-time validation; potential overfitting.
Arshad et al. [6]	Tuned MLP and classical ML models (SVM, RF)	No XAI; dataset dependency; no external validation; limited clinical applicability.
Swain [11]	k-NN, SVM, LDA, CNN comparison on UCI dataset	No XAI; data leakage risk; small dataset; lacks diversity; no deep interpretability.
Tougui et al. [16]	Wav2Vec2 and HuBERT pretrained ASR models on French smartphone voice data	High computational cost; single-language bias; self-reported data; limited generalization; no XAI.
Tekindor et al. [26]	Convolutional Autoencoder + LSTM hybrid	No XAI; small dataset; potential overfitting; high computational load.
Xu et al. [27]	Random Forest and Gradient Boosting with SHAP feature ranking	Limited dataset diversity; no spontaneous speech; lacks multimodal validation.
Lim et al. [20]	Random Forest and AdaBoost using acoustic and linguistic features	No XAI; cross-lingual bias; small training set; no medication metadata.
Malekroodi et al. [21]	Wav2Vec2 and HuBERT with Supervised Contrastive Loss (SupCon)	High computational cost; no XAI; single-language data; absence of clinical metadata.
Our Study (2025)	Optimized XGBoost	Contributions: implemented recursive feature elimination (RFE) feature optimization; grid search hyperparameter tuning; SHAP-based explainability; low computational cost; sensitivity analysis; interpretable and generalizable PD voice detection framework.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis has investigated the use of voice analysis and sophisticated machine learning techniques for an early indication of Parkinson's disease. The research was inspired by the understanding that Parkinson's disease is still a significant problem in today's healthcare, and diagnosing the condition can be delayed based on subjective clinical evaluations and expensive imaging methods. Given that voice impairments are often present at early stages of the disease, vocal signal can potentially function as an accessible and low-cost way to serve as a non-invasive biomarker.

Via systematic pre-processing, feature selection and classification, a framework was proposed for the discrimination of Parkinson's Disease patients from normal subjects. The data quality was guaranteed by preprocessing the given dataset to avoid redundancy, outlier, and imbalance. Dimensionality reduction and retained discriminative information by feature engineering through statistical tests, SelectKBest ranking and principal component analysis. Several machine-learning classifiers, including decision tree, ensemble methods and gradient boosting machines, were fitted using cross-validation and optimized by hyperparameter tuning. Of these, XGBoost with 100 features ranked highest based on all five most common evaluation metrics: accuracy, precision, recall, F1-score and Cohen's Kappa. Reliability of the probabilistic outputs was also supported by calibration measures (e.g., Brier score). Furthermore, interpretable AI methods, such as SHAP and Morris sensitivity analysis allowed for interpretability by explaining how the acoustic features affected classification. These results emphasized jitter and shimmer, the MFCC coefficients, as well as some non-linear wavelet-based measures as informative descriptors for Parkinson's-related speech alteration further supporting the clinical relevance of the computational models.

The work also contributes significantly by showing that a pipeline composed of preprocessing, dimensionality reduction, robust ensemble learning and explainability is crucial for increasing classification performance in the Parkinson's disease voice. It also demonstrates how computational strategies could be connected to biomedical experience as a bridge the gap between data-driven deducing and clinical use. These results suggest that voice, coupled with interpretable machine learning, represent a promising candidate to be adopted as an accurate digital biomarker for early diagnosis and remote monitoring in clinical and telemedicine settings.

5.2 Future Work

This work has produced some promising results; however there are a number of ways in which it may be extended to broaden its scope and impact. Perhaps the most critical problems are small and biased datasets. In future work, the recording and analysis of large-scale, heterogeneous and multilingual TDVCs in which variability can be captured from a wide range of speaker populations (age groups, dialects) as well as acoustic conditions should be placed into focus. Extending to non-vowels, conversational speech, reading tasks and spontaneous dialogues would improve the generalisation of models in real-life situations. The addition of longitudinal datasets for monitoring disease progression over time (important for clinical decision support) would also be possible.

Third, the development of novel deep and hybrid models is another significant direction. Although ensemble-based methods such as XGBoost have achieved excellent performances, deep learning models, including convolutional and transformer-based networks, may provide further performance improvements for their training on sufficiently large datasets. Hybrid methods, that leverage the hand crafted acoustic features with learned embeddings for overall predictability and interpretability could get a way lead to bridge this gap. Meanwhile, methods like model quantization and pruning should be investigated to enable deep models to be deployed on mobile and edge devices efficiently for clinical deployment at scale.

Another exciting area is multimodal data fusion. Voice could also be overlaid with gait, facial expression, clinical metadata (e.g., medication history and disease duration), or handwriting samples to create more robust diagnostic models. Such multimodal methods could increase robustness and be able to identify aspects of Parkinson's disease, which are not entirely available in speech alone. This integration, however, also drives the demand for more advanced fusion strategies and interpretability frameworks to make sure models remain interpretable for clinicians.

Finally, regulatory ethical and clinical translation concerns need to be overcome. Although interpretable AI techniques are presented in this thesis, more efforts are needed development standards for interpretability and to validate in clinical trials. Privacy, consent, and bias While algorithms help to automate the analysis of vocal biomarkers in disease diagnostics, privacy advocacy organisations have raised concerns about how voice data is being collected and used. Overcoming such challenges will enable future research to extend the groundwork laid herein, and provide more robust, pragmatic and ethically acceptable tools for early diagnosis and monitoring of Parkinson's disease.

Bibliography

[1] P. Mallela et al., "Dysarthria type specific detection of neurological disorders from speech," in *Proc. INTERSPEECH 2020*, 2020, pp. 2273–2277.

[2] C. H. Adler et al., "Clinical diagnostic accuracy of early/advanced parkinson disease: An updated clinicopathologic study," *Neurology. Clinical practice*, vol. 11, no. 4, e414–e421, 2021. DOI: 10.1212/CPJ.0000000000001016. [Online]. Available: https://doi.org/10.1212/CPJ.0000000000001016.

[3] Y. Quamar, W. Jan, A. Khan, Z. Mehmood, R. Ezemain, and M. I. Razzak, "Voice-based early diagnosis of parkinson's disease using spectrogram and artificial intelligence models," *Journal of Medical Systems*, vol. 46, 2022. DOI: 10.1007/s10916-022-01861-8.

[4] A. Suppa et al., "Voice in parkinson's disease: A machine learning study," *Frontiers in Neurology*, vol. 13, p. 831 428, 2022. DOI: 10.3389/fneur.2022.831428. [Online]. Available: https://doi.org/10.3389/fneur.2022.831428.

[5] S. B. Appakaya, R. Pratihar, and R. Sankar, "Parkinson's disease classification framework using vocal dynamics in connected speech," *Algorithms*, vol. 16, no. 11, p. 1575, 2023. DOI: 10.3390/a16111575.

[6] U. Arshad, J. U. Malik, et al., "Machine learning approaches to identify parkinson's disease using voice signal features," *Frontiers in Artificial Intelligence*, 2023. DOI: 10.3389/frai.2023.1698417.

[7] A. Rehman et al., "Parkinson's disease detection using hybrid lstm-gru deep learning model," *Sustainability*, vol. 15, no. 3, 2023. DOI: 10.3390/su15032598.

[8] T. Saba, N. Alharbi, M. A. El-Aziz, and I. Ullah, "Parkinson's disease detection using hybrid lstm–gru deep learning model," *Electronics*, vol. 12, no. 18, 2023. DOI: 10.3390/electronics12184224.

[9] S. Scimeca et al., "Robust and language-independent acoustic features in parkinson's disease," *Frontiers in Neurology*, vol. 14, 2023. DOI: 10.3389/fneur.2023.1198058.

[10] V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-modal deep learning diagnosis of parkinson's disease—a systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2399–2423, 2023. DOI: 10.1109/TNSRE.2023.3277749.

[11] K. Swain, "Towards early intervention: Detecting parkinson's disease using voice disorders via machine learning," *Open Biomedical Engineering Journal*, vol. 17, pp. 15–21, 2023. DOI: 10.2174/1874120702317010015.

[12] S. N. H. Bukhari and K. A. Ogudo, "Ensemble machine learning approach for parkinson's disease detection using speech signals," *Mathematics*, vol. 12, no. 10, 2024. DOI: 10.3390/math12101575.

[13] K. R. Chaudhuri et al., "Economic burden of parkinson's disease: A multinational, real-world, cost-of-illness study," *Drugs - Real World Outcomes*, vol. 11, no. 1, pp. 1–11, 2024. DOI: 10.1007/s40801-023-00410-1. [Online]. Available: https://doi.org/10.1007/s40801-023-00410-1.

[14] C. R. Dhivyaa, K. Nithya, and S. Anbukkarasi, "Enhancing parkinson's disease detection and diagnosis: A survey of integrative approaches across diverse modalities," *IEEE Access*, vol. 12, pp. 158 999–159 024, 2024. DOI: 10.1109/ACCESS.2024.3487001.

[15] R. R. Ilesan et al., "In silico decoding of parkinson's: Speech writing analysis," *Journal of Clinical Medicine*, vol. 13, no. 18, 2024. DOI: 10.3390/jcm13185573.

[16] I. Tougui, K. Mijwil, et al., "Federated occlusion learning for parkinson's disease detection from acoustic voice features," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 2529–2536, 2024. DOI: 10.1109/JBHI.2023.3273310.

[17] A. Adnan, K. Y. Chan, et al., "A novel fusion architecture for detecting parkinson's disease using semi-supervised speech embeddings," *NPJ Digital Medicine*, vol. 8, 2025. DOI: 10.1038/s41746-025-XXXX-X.

[18] B. Egbo, Z. Nigmetolla, N. A. Khan, and P. K. Jamwal, "Explainable machine learning for early detection of parkinson's disease in aging populations using vocal biomarkers," *Frontiers in Aging Neuroscience*, vol. 17, 2025. DOI: 10.3389/fnagi.2025.1672971.

[19] M. D. Khedimi, E. Gazit, T. Herman, et al., "A unified deep learning ensemble framework for voice-based parkinson's disease detection and severity prediction," *Frontiers in Aging Neuroscience*, vol. 17, 2025. DOI: 10.3389/fnagi.2025.1092426.

[20] W.-S. Lim et al., "A cross-language speech model for detection of parkinson's disease," *Journal of Neural Transmission*, vol. 132, pp. 579–590, 2025. DOI: 10.1007/s00702-024-02649-7.

[21] E. Malekroodi, F. Bayram, S. Kamarthi, T. Powell, and P. Jamwal, "Speech-based parkinson's detection using pre-trained self-supervised asr models and supervised contrastive learning," *Bioengineering*, vol. 12, no. 7, p. 728, 2025. DOI: 10.3390/bioengineering12070728.

[22] I. Naeem et al., "Voice biomarkers as prognostic indicators for parkinson's disease using machine learning," *Scientific Reports*, vol. 15, 2025. DOI: 10.1038/s41598-025-96575-9.

[23] Y. Rahmatallah, A. Iyer, et al., "Explainable artificial intelligence to diagnose early parkinson's disease via voice analysis," *Scientific Reports*, vol. 15, 2025. DOI: 10.1038/s41598-025-25415-1.

[24] M. Rashidi, S. Arima, et al., "Prediction of parkinson disease using long-term, short-term acoustic features based on machine learning," *Brain Sciences*, vol. 15, no. 7, 2025. DOI: 10.3390/brainsci15070739.

[25] M. Shen, P. Mortezaagha, and A. Rahgozar, "Explainable machine learning for early detection of parkinson's disease in aging populations using vocal biomarkers," *Scientific Reports*, vol. 15, 2025. DOI: 10.1038/s41598-025-96575-9.

[26] H. Tekindor, A. Kilinc, et al., "Speech signals-based parkinson's disease diagnosis using deep learning (hybrid autoencoder–lstm)," *Computers in Biology and Medicine*, vol. 152, 2025. DOI: 10.1016/j.compbiomed.2022.106341.

[27] R. Xu et al., "Non-invasive detection of parkinson's disease based on speech analysis and interpretable machine learning," *Frontiers in Aging Neuroscience*, vol. 17, 2025. DOI: 10.3389/fnagi.2025.1190788.