4. Conclusion

Author: Alex Terzis

Abstract

This paper investigates the performance of multiple machine learning models for detecting Parkinson's Disease using voice features, including both classical and information-theoretic features (e.g., RPDE and DFA). A total of 8 models were compared: Logistic Regression, SVM, KNN, Random Forest, Gradient Boosting, Decision Tree, Naive Bayes, and QDA. Metrics such as accuracy, ROC AUC, confusion matrices, feature importances, runtime benchmarks, and t-SNE visualizations were analyzed.

1. Introduction

Parkinson's Disease (PD) is a neurodegenerative disorder characterized by motor and non-motor symptoms. Voice impairment is an early biomarker for PD, making it a non-invasive and cost-effective diagnostic tool. In this study, we utilize the well-known voice dataset from the UCI repository and evaluate the discriminative power of different feature sets and classifiers.

2. Methods

We use a set of voice features including fundamental frequency, jitter, shimmer, noise-to-harmonic ratio (NHR), nonlinear dynamic complexity (DFA), and signal entropy (RPDE). Models were evaluated using stratified cross-validation. GridSearchCV was used for hyperparameter tuning where applicable. Visual diagnostics included ROC curves, confusion matrix breakdowns, t-SNE clustering, and runtime comparisons.

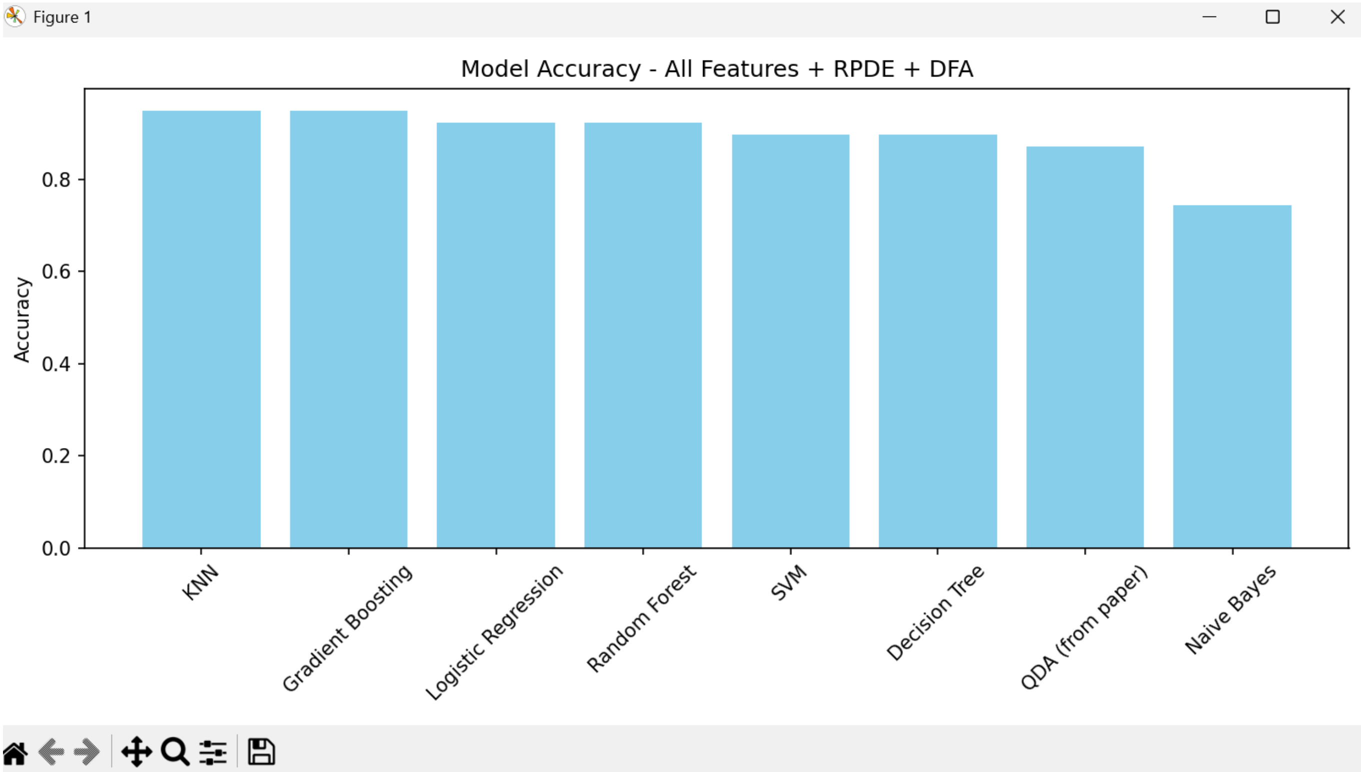
3. Results

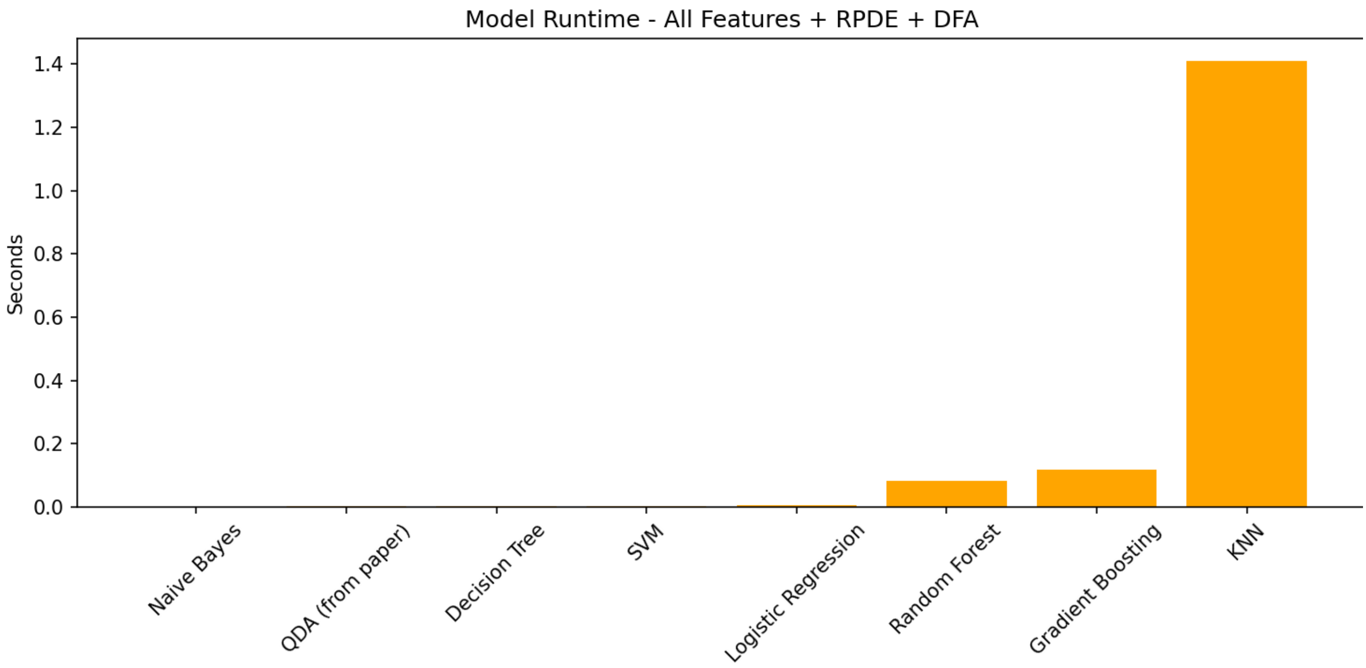
KNN, Gradient Boosting, and Logistic Regression achieved the highest accuracies (~95%) using the full feature set. Random Forest and Gradient Boosting were most accurate with RPDE+DFA only. Naive Bayes performed weakest. Runtime varied substantially, with KNN being the slowest. Feature importance analysis revealed RPDE and PPE as top predictors. t-SNE clustering further confirmed class separability for the best model (KNN).

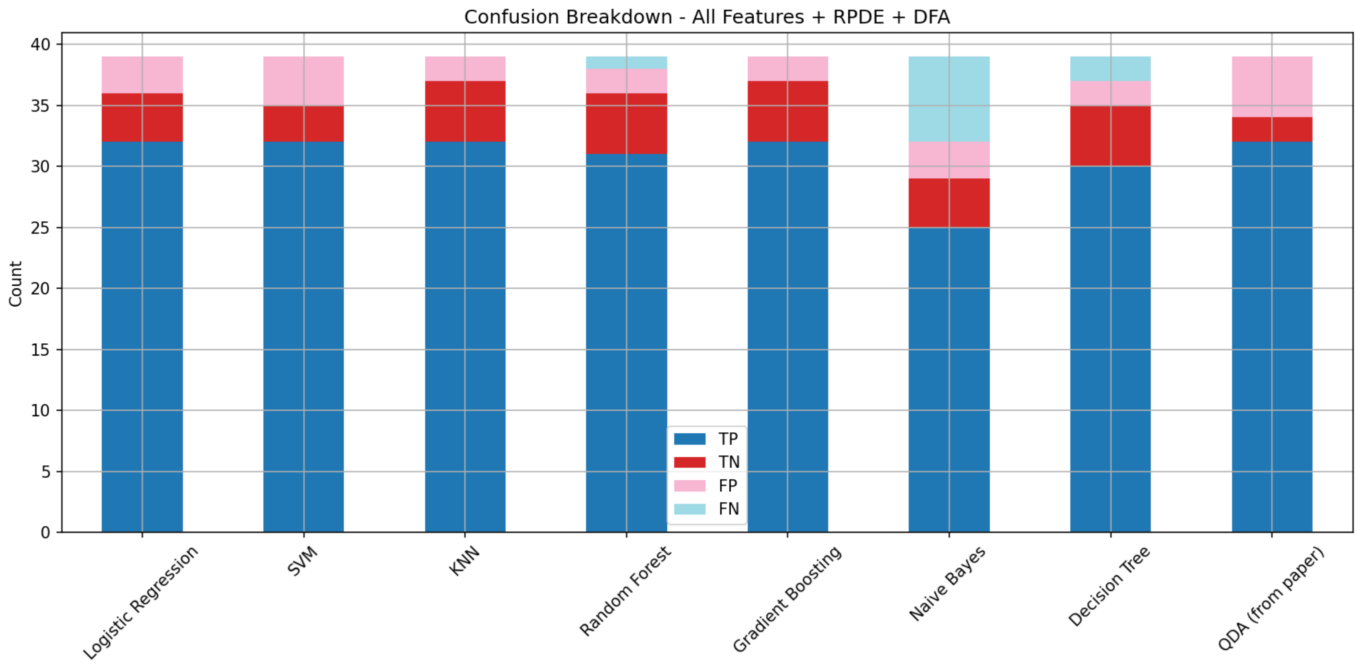
4. Conclusion

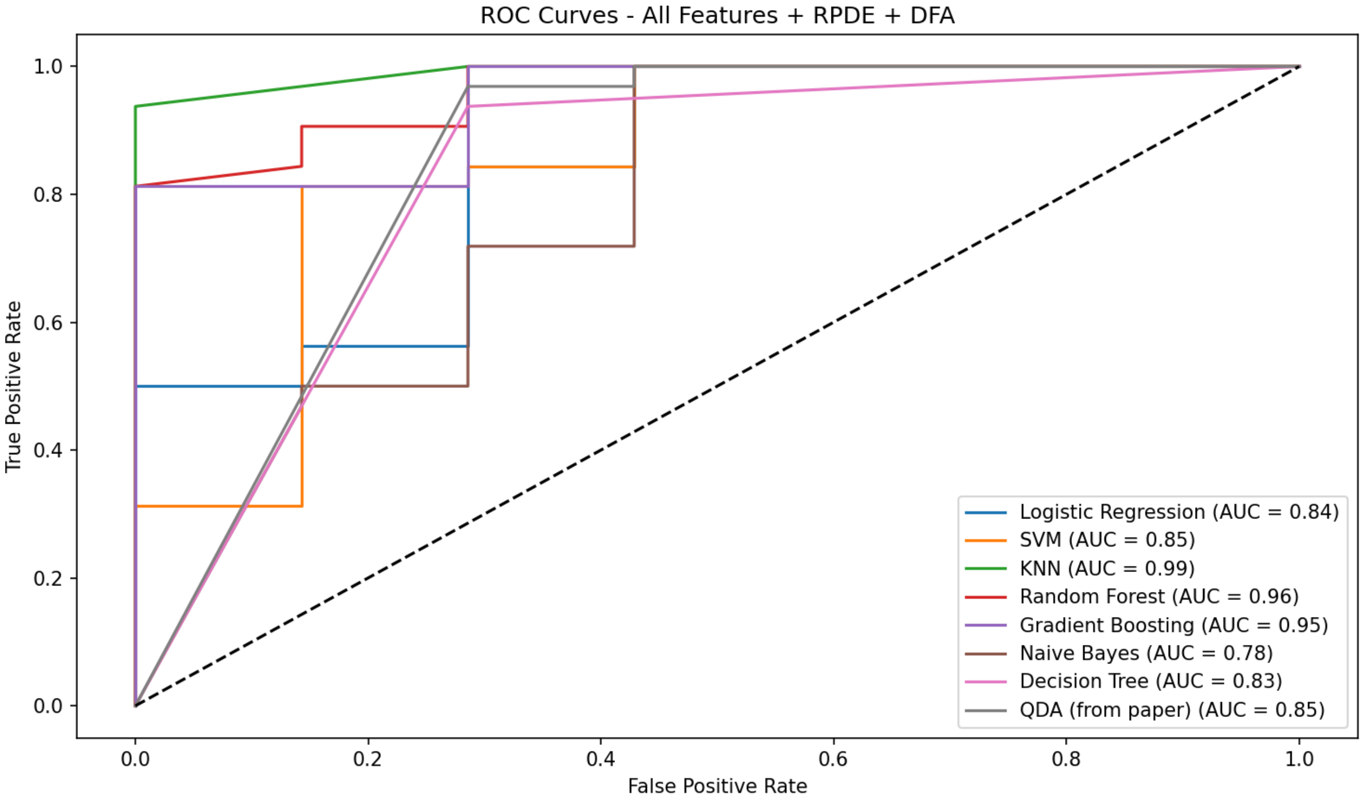
Results show that modern ensemble methods and KNN are highly effective for PD voice detection when all features are included. However, even just RPDE and DFA maintain strong performance. Future work includes deploying this in mobile apps using TensorFlow Lite, and expanding the dataset to include more diverse patient recordings.

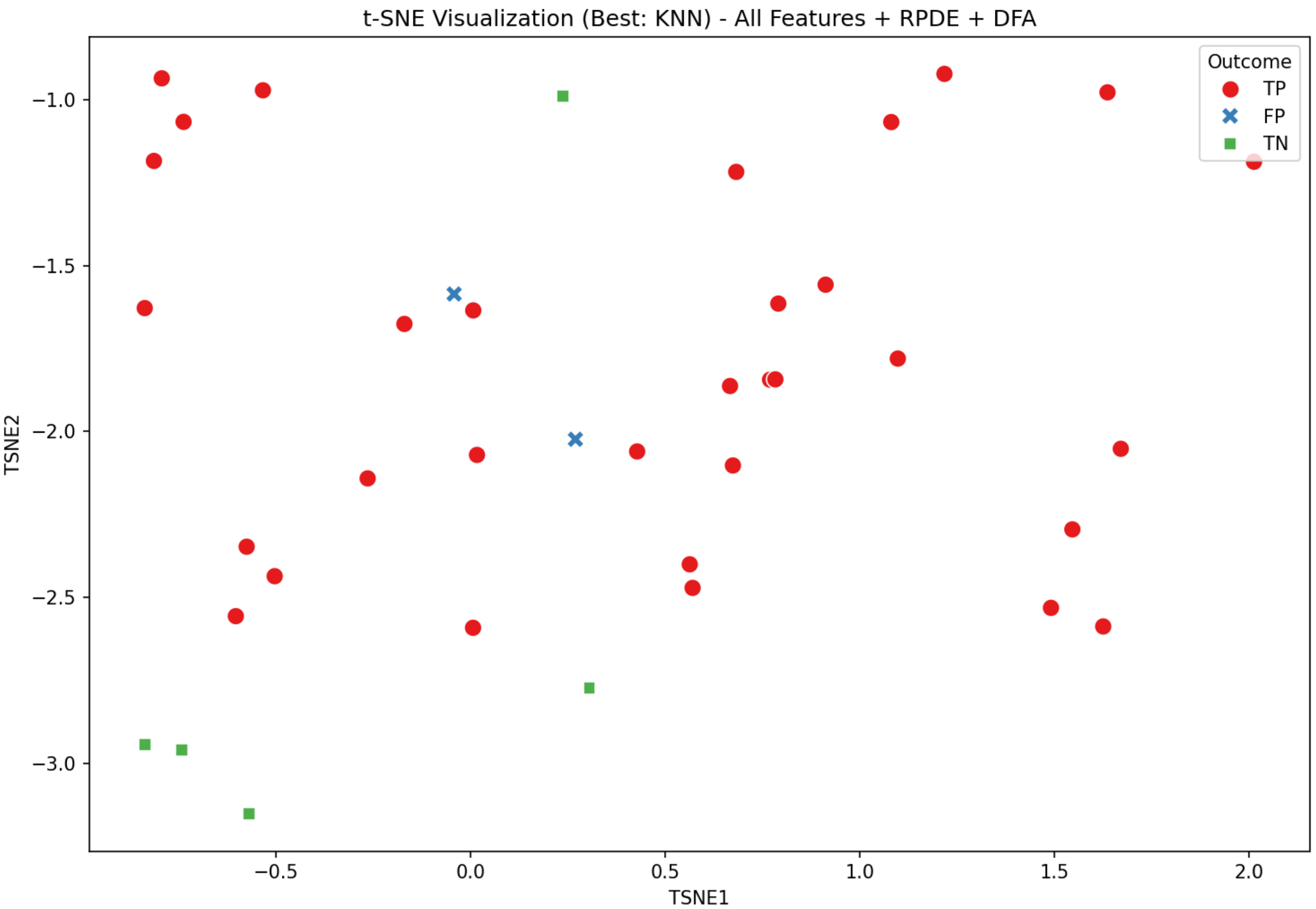
5. Visual Results

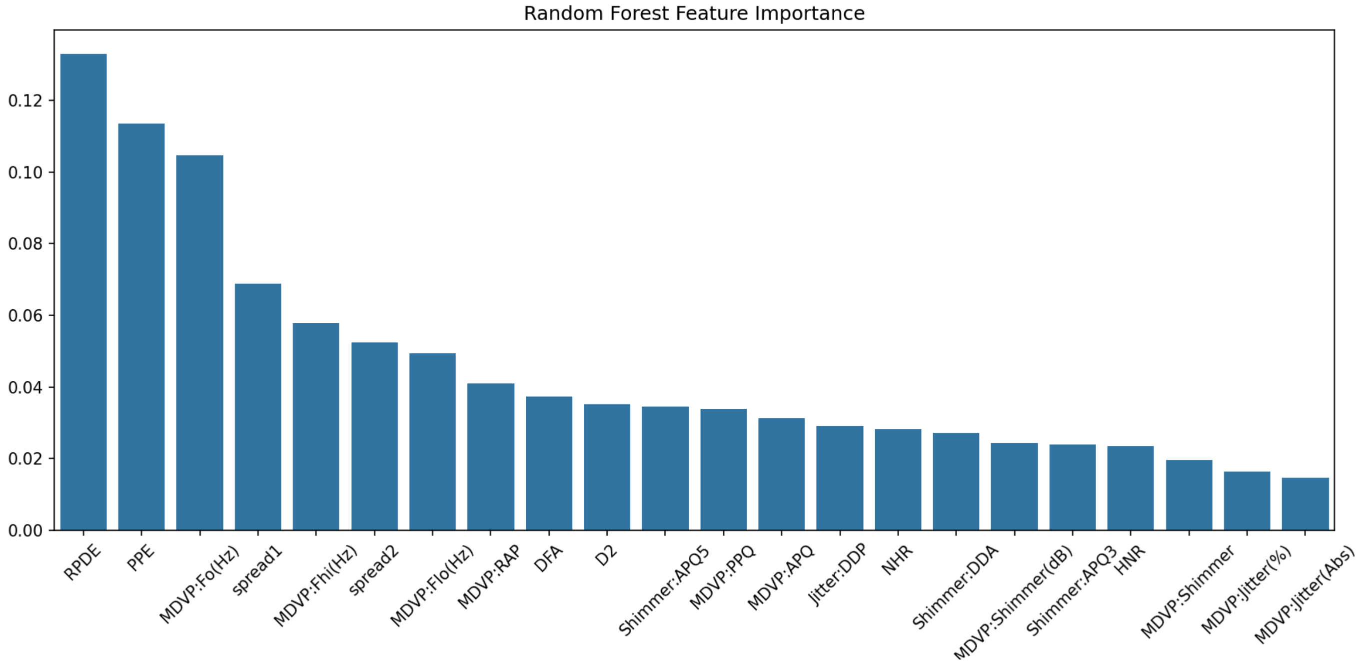


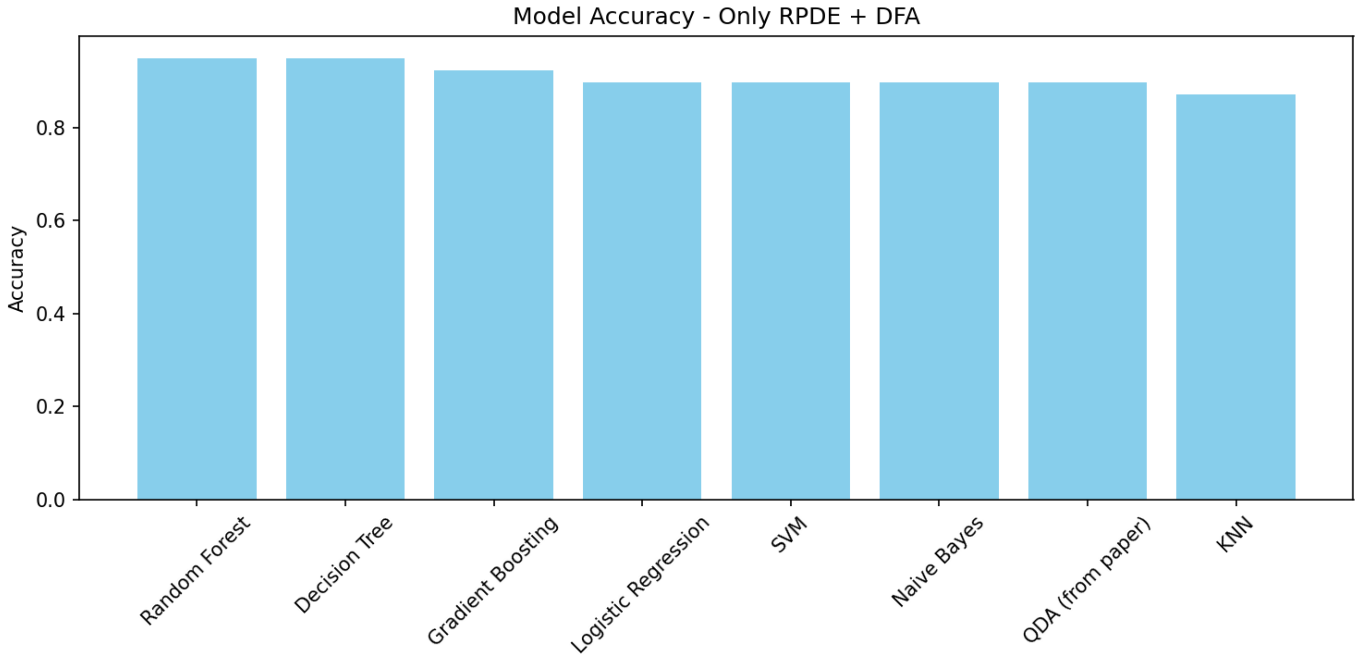

Model Runtime - All Features + RPDE + DFA
Seconds
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Naive Bayes
QDA (from paper)
Decision Tree
SVM
Logistic Regression
Random Forest
Gradient Boosting
KNN

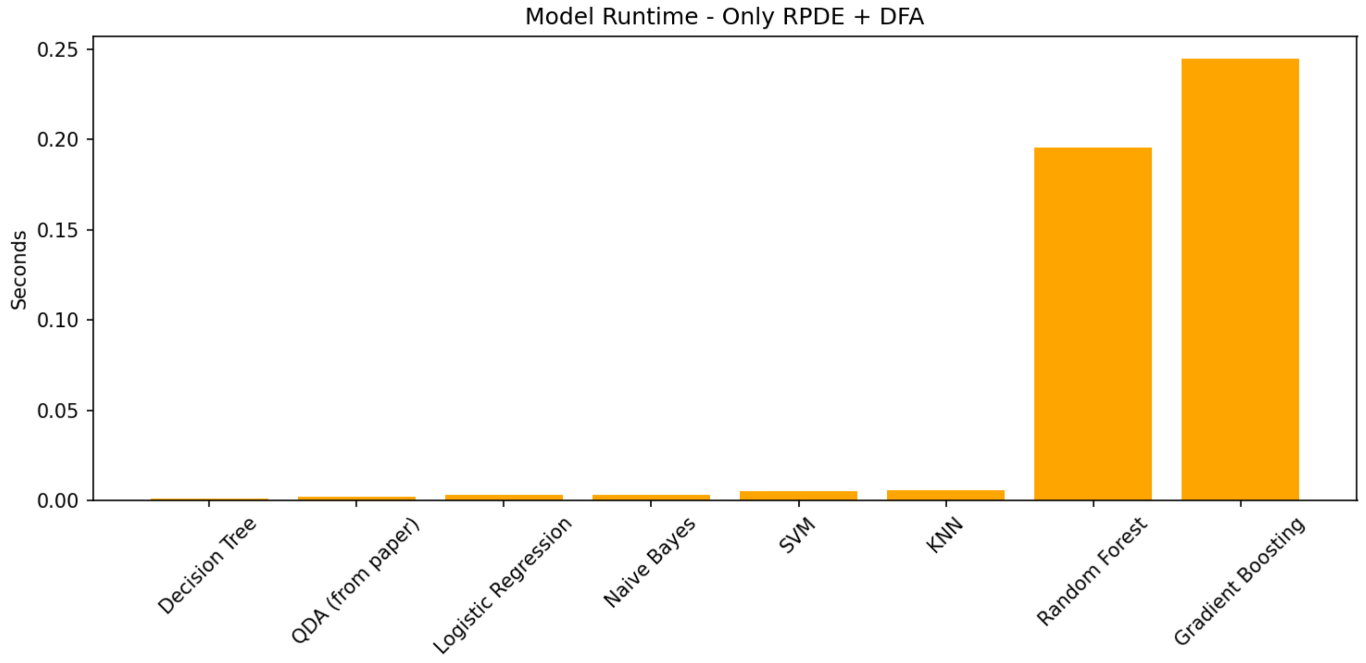

Confusion Breakdown - All Features + RPDE + DFA
Count
40
35
30
25
20
15
10
5
0
TP
TN
FP
FN
Logistic Regression
SVM
KNN
Random Forest
Gradient Boosting
Naive Bayes
Decision Tree
QDA (from paper)

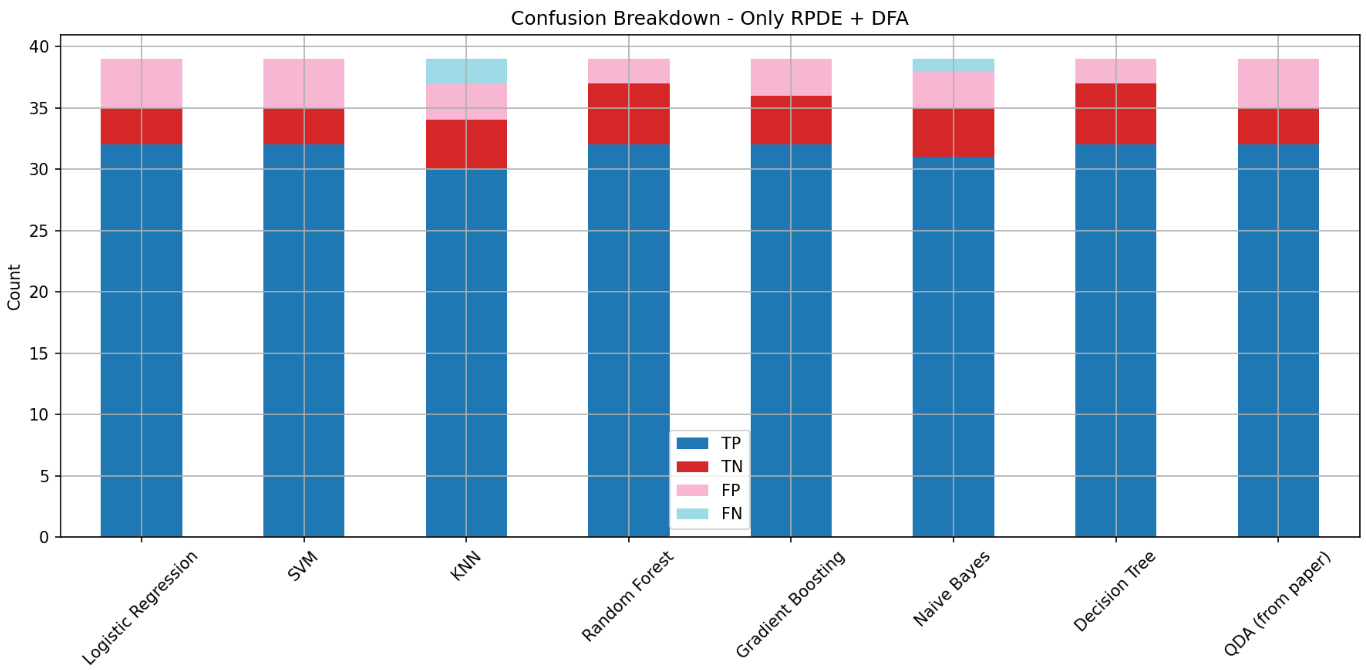

ROC Curves - All Features + RPDE + DFA
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.84)
SVM (AUC = 0.85)
KNN (AUC = 0.99)
Random Forest (AUC = 0.96)
Gradient Boosting (AUC = 0.95)
Naive Bayes (AUC = 0.78)
Decision Tree (AUC = 0.83)
QDA (from paper) (AUC = 0.85)

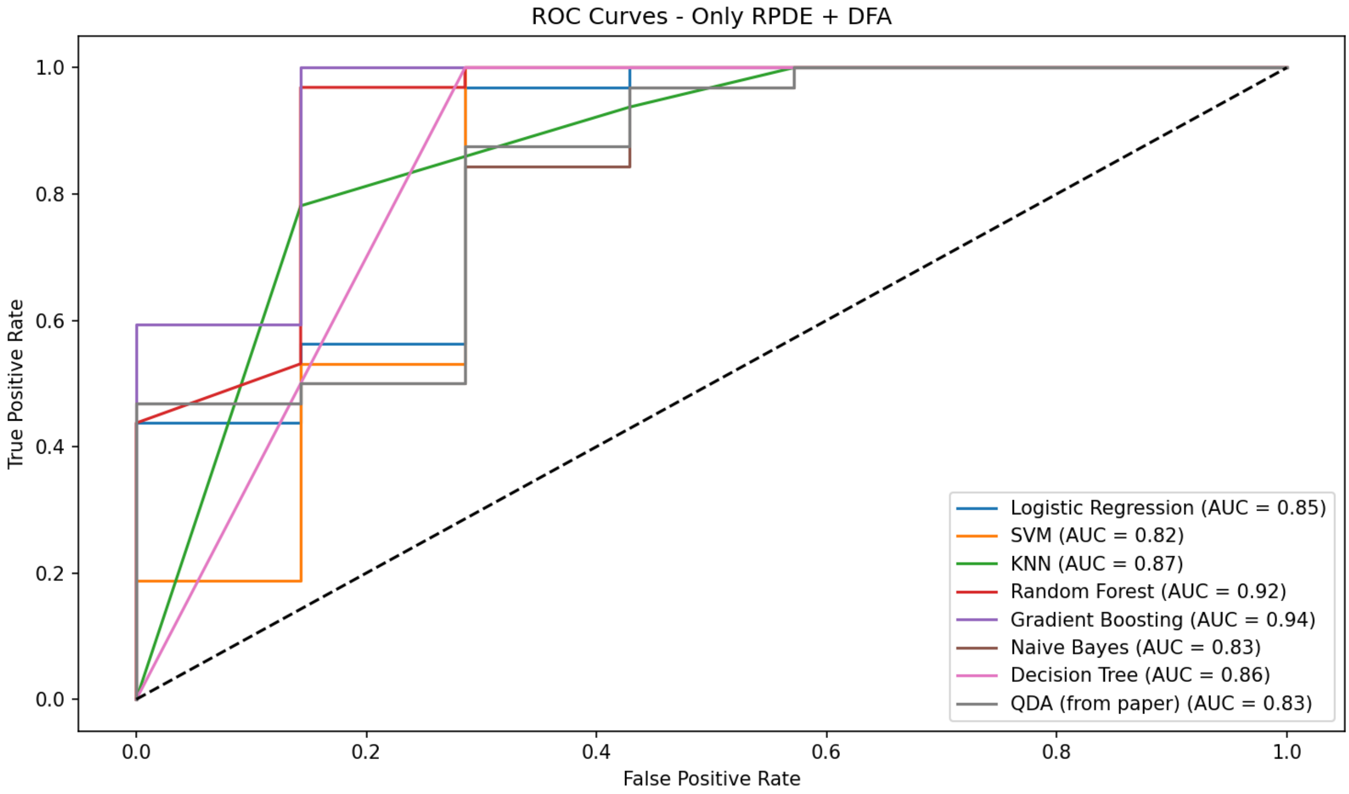

t-SNE Visualization (Best: KNN) - All Features + RPDE + DFA
TSNE2
TSNE1
−1.0
−1.5
−2.0
−2.5
−3.0
−0.5
0.0
0.5
1.0
1.5
2.0
Outcome
TP
FP
TN


Random Forest Feature Importance
0.12
0.10
0.08
0.06
0.04
0.02
0.00
RPDE
PPE
MDVP:Fo(Hz)
spread1
MDVP:Fhi(Hz)
spread2
MDVP:Flo(Hz)
MDVP:RAP
DFA
D2
Shimmer:APQ5
MDVP:PPQ
MDVP:APQ
Jitter:DDP
NHR
Shimmer:DDA
MDVP:Shimmer(dB)
Shimmer:APQ3
HNR
MDVP:Shimmer
MDVP:Jitter(%)
MDVP:Jitter(Abs)


Model Accuracy - Only RPDE + DFA
Accuracy
0.0
0.2
0.4
0.6
0.8
Random Forest
Decision Tree
Gradient Boosting
Logistic Regression
SVM
Naive Bayes
QDA (from paper)
KNN


Model Runtime - Only RPDE + DFA
Seconds
0.25
0.20
0.15
0.10
0.05
0.00
Decision Tree
QDA (from paper)
Logistic Regression
Naive Bayes
SVM
KNN
Random Forest
Gradient Boosting


Confusion Breakdown - Only RPDE + DFA
Count
40
35
30
25
20
15
10
5
0
TP
TN
FP
FN
Logistic Regression
SVM
KNN
Random Forest
Gradient Boosting
Naive Bayes
Decision Tree
QDA (from paper)


ROC Curves - Only RPDE + DFA
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.85)
SVM (AUC = 0.82)
KNN (AUC = 0.87)
Random Forest (AUC = 0.92)
Gradient Boosting (AUC = 0.94)
Naive Bayes (AUC = 0.83)
Decision Tree (AUC = 0.86)
QDA (from paper) (AUC = 0.83)