Digital Signal Processing and Machine Learning in Voice Analysis

By Choudhary Sumit Jaswant (22/11/EE/010), Vaibhav Mishra (22/11/EE/027), Siddharth Kumar (22/11/EE/028), Aman Kushwaha (22/11/EE/053)

School Of Engineering, Jawaharlal Nehru University

Abstract: In the following project, a voice signal was analyzed based on some DSP and ML algorithms to judge the performance comparison of different classification models. At the first step, the voice signal is recorded in .wav format, followed by adding artificial noise and phase distortion to generate signals simulating real-case scenarios. The distorted voice signal was decomposed using Fourier decomposition to extract the features. The extracted features were used as inputs in various machine learning classifiers, which included LSTM RNN, Support Vector Machines (SVM), Decision Trees, and Random Forests. The performance of the models was checked and compared in terms of the accuracy of the classification of the distorted voice features. Cross-validation was employed to ensure the robustness of the models. This study provided insight into how machine learning techniques can be applied to voice signal processing.

Introduction: Voice signal processing is crucial for applications such as speech recognition, noise filtering, and voice-based security systems. This project investigates the ability of various machine learning models in classifying voice signals distorted by noise and phase shifts. In this process, feature extraction, which is a popular approach in DSP, decomposes a signal into its constituent parts in the frequency domain. These extracted features were then used to train and test various machine learning models to assess their classification performance.

Methodology:

Voice Signal Acquisition: A voice signal was recorded in .wav format. This is the basic input which will further undergo changes.

Signal Generation: The following types of signals were generated:

- Clean signals were duplicated from the original voice signal.
- Noisy signals were created by adding Gaussian noise to the clean signals.

- Phase-shifted signals were generated by introducing random phase shifts to the clean signals.

These signals were then labeled as 0 (clean), 1 (noisy), and 2 (phase-shifted), forming a labeled dataset for model training and evaluation.

Fourier Decomposition: Fourier decomposition was used to transform the disturbed voice signals into their frequency domain equivalents by converting time-domain signals into the frequency domain. Features were subsequently extracted in the frequency domain, which played a central role in the classification.

Machine Learning Models: The extracted features were fed into a variety of machine learning models for classification:

- LSTM RNN: A deep learning model for sequence data.
- Support Vector Machines (SVM)
- Decision Trees
- Random Forest

Random Forest Dimensionality reduction and standardization:

The features were standardized using StandardScaler to ensure that all features had zero mean and unit variance.

Using Principal Component Analysis (PCA), reduced the number of features to 10 components.

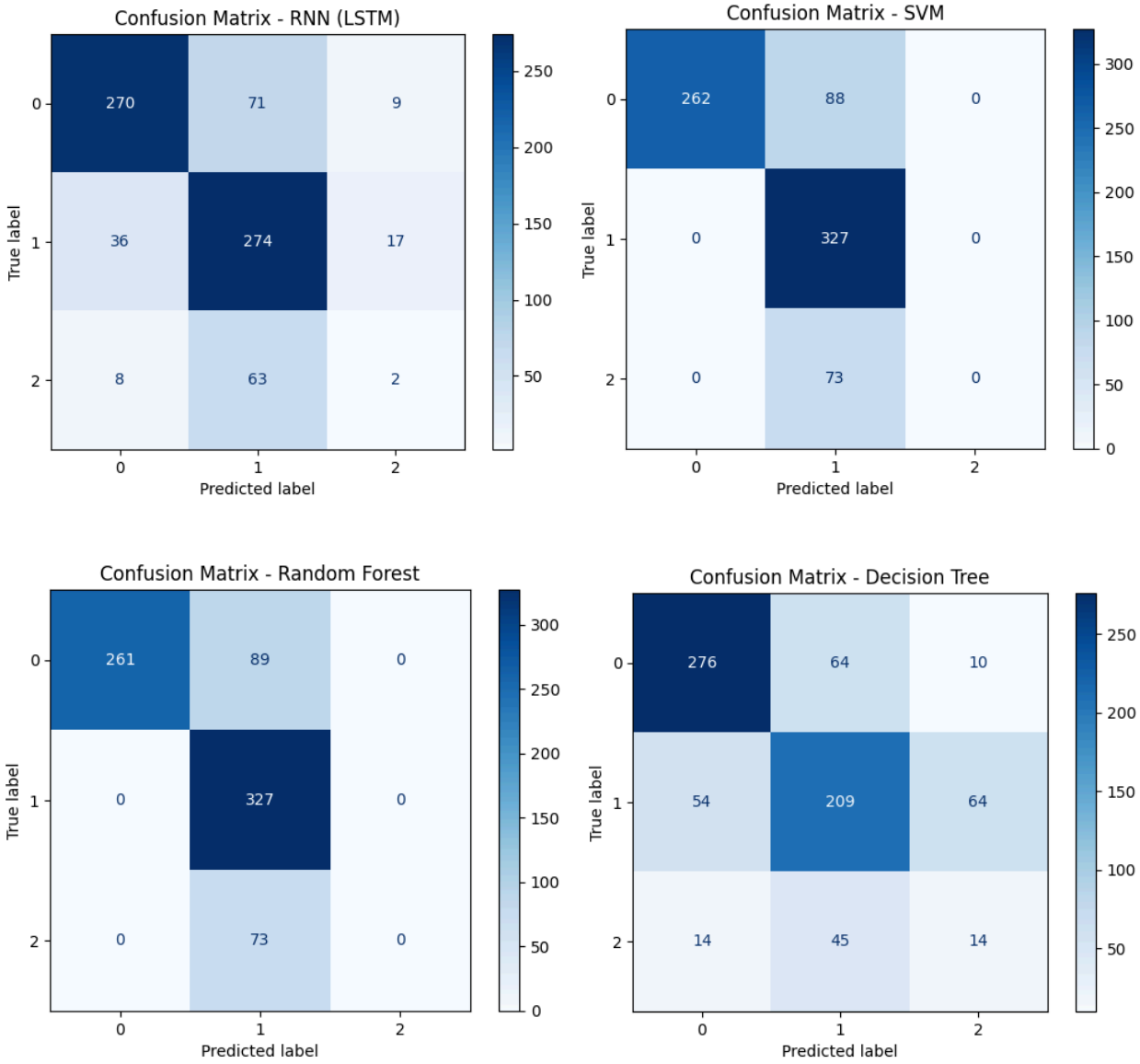
Model Training and Testing:

- **Random Forest:** Hyperparameter tuning for Random Forest was performed using **RandomizedSearchCV**, exploring parameters such as the number of estimators, maximum depth, and minimum samples for splits and leaves.
- **Cross-Validation:** Cross-validation (Stratified K-fold) was employed for all models to evaluate their performance on different subsets of the data, ensuring a robust evaluation of their classification ability.

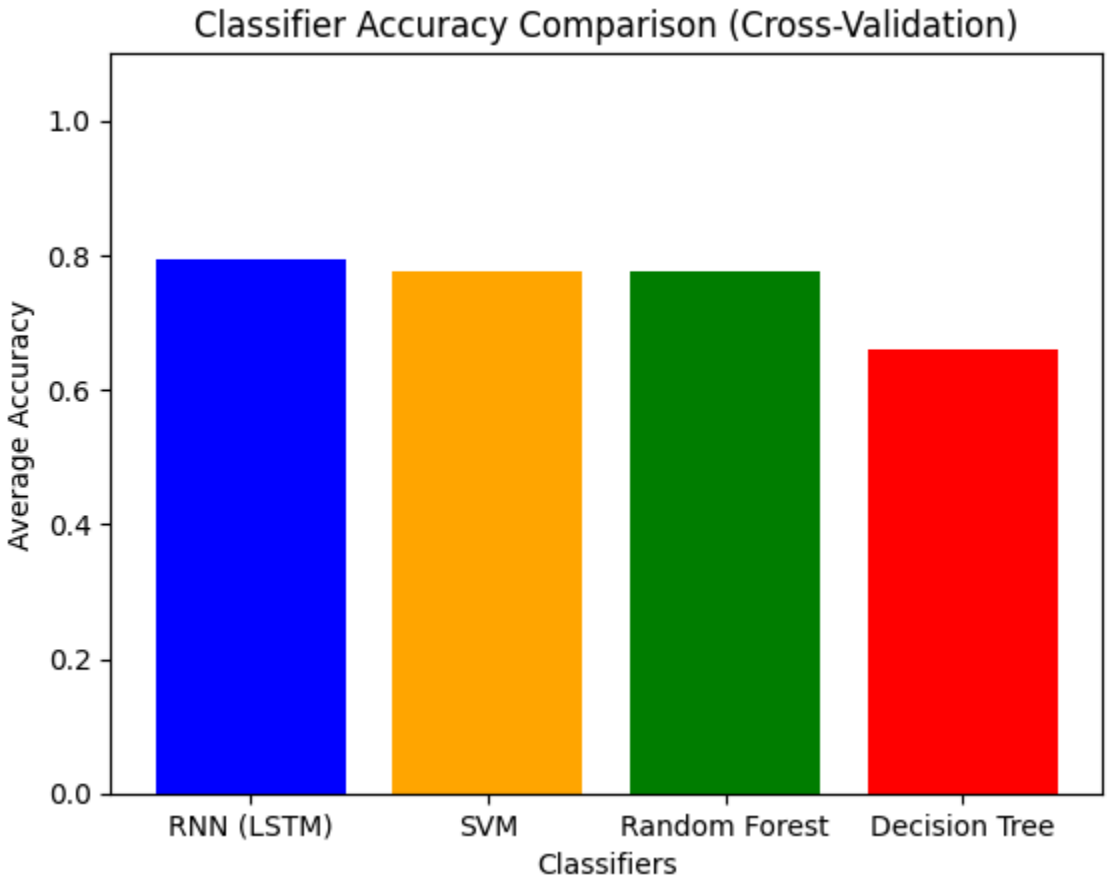
Evaluation:
Each model was trained using the processed features and assessed for accuracy. For visualization, classification performance plots and confusion matrices for all the classifiers were generated.

Results:

```
RNN (LSTM) Average Cross-Validation Accuracy: 0.7938
SVM Average Cross-Validation Accuracy: 0.7751
Random Forest Average Cross-Validation Accuracy: 0.7758
Decision Tree Average Cross-Validation Accuracy: 0.6613
```



Label 0 represents clean signal, Label 1 represents noisy signal & Label 2 represents phase shifted signal



The machine learning models were trained on the Fourier features, and their performance was evaluated using accuracy. The confusion matrix for all classifiers was plotted, and the accuracy of each model was compared. Cross-validation results were used to gauge model robustness, with LSTM RNN showing strong performance in handling sequential data patterns.

Conclusion:

This project demonstrated the effectiveness of using Fourier decomposition for feature extraction from voice signals affected by noise and phase distortions. The machine learning models, including **LSTM RNN**, **SVM**, **Decision Trees**, and **Random Forest**, were tested for their performance in classifying distorted voice features. The results provided valuable insights into the suitability of these models for various signal processing tasks and their robustness in handling real-world distortions in voice signals. LSTM RNN, in particular, exhibited strong performance in capturing sequential patterns in the data, while other models such as SVM and Random Forest showed reliable classification accuracy. Cross-validation ensured that the models' performances were robust and generalized well to different subsets of the data. Overall, this study highlights the potential of combining DSP techniques with advanced machine learning models for improving the accuracy of voice signal classification in practical applications.

Source Code:

https://github.com/siddharthk7704/DSP-FDM

Source Code of Fourier Decomposition Method used:

https://github.com/udawat/Fourier-Decomposition-Method