18-752 Project Report

Note Recognition in Renditions of Piano Instruments

Li-Wei Chi

Leron Julian

1. **Project Scope:**

   The scope of this project is to train a note recognition model using piano pieces found online. This model will classify notes between C4 (262 Hz) and C6 (1040 Hz), 24 notes in total, usually played by right hand. To collect data for this task, we download 10 piano pieces online [1] in the form of MIDI files in which we have all the notes in a piano piece (y) and the time at which these notes are played; the onsets. To extract the notes being played in the audio (WAV) file, we extract the nearby sound waves corresponding to each note (X). Our data collection process was opposite to the common data collection process ($X \rightarrow y$). Instead, we did $y \rightarrow X$. We then take the mean across the spectrogram of the note played. To label the notes, we framed the problem into 24 binary classifications (i.e. is C5 played? Is F5 played? …), so our target is a 24-dimensional multi-hot vector. Visualization of the data in the form of a spectrogram as well as the distribution of notes in the train, validation, and test sets are included in the appendix.

2. **Feature Extraction and Exploratory Data Analysis (EDA):**

   To get a better understanding of our data, we visualize the audio waveforms using the librosa library. To analyze the frequency components, we turned the audio signals into spectrograms. Upon inspecting the spectrogram, we noticed that the primary frequency as well as the harmonic frequencies of notes show up when played, this makes the classification problem more challenging since notes that are an octave apart would have similar frequency components and harder to tell apart by human eye. Since the spectrogram is a very high dimensional feature space, we performed steps to reduce the dimensionality of the data. First, we take the mean of the spectrogram matrix across time, which yields a 4096 dimensional vector, which is the average power at each frequency component over time. Then, we experimented with 3 feature extraction methods: Linear Discriminant Analysis (LDA), Latent Semantic Analysis (LSA), and Independent Component Analysis (ICA).

Since recognizing if a note is being played is a binary classification problem. We initially use LDA feature extraction with targets being whether each note is played, projecting samples to 1-dimension. The LDA figure in the appendix below shows a visualization of this. We color the points, based on whether the note was played in the sample (blue when it is played and red when it is not). The purplish points in the middle, shows overlapping of red and blue samples (so the data is not completely separable in 1D). We also present visuals of LSA and ICA in the appendix below. Moving forward to our classification methods, we use LDA as our primary feature extraction method. To also compare the results in terms of classification accuracy, we experiment with using LSA as a feature extraction method and compare it to LDA.

3. **Classification Methods:**

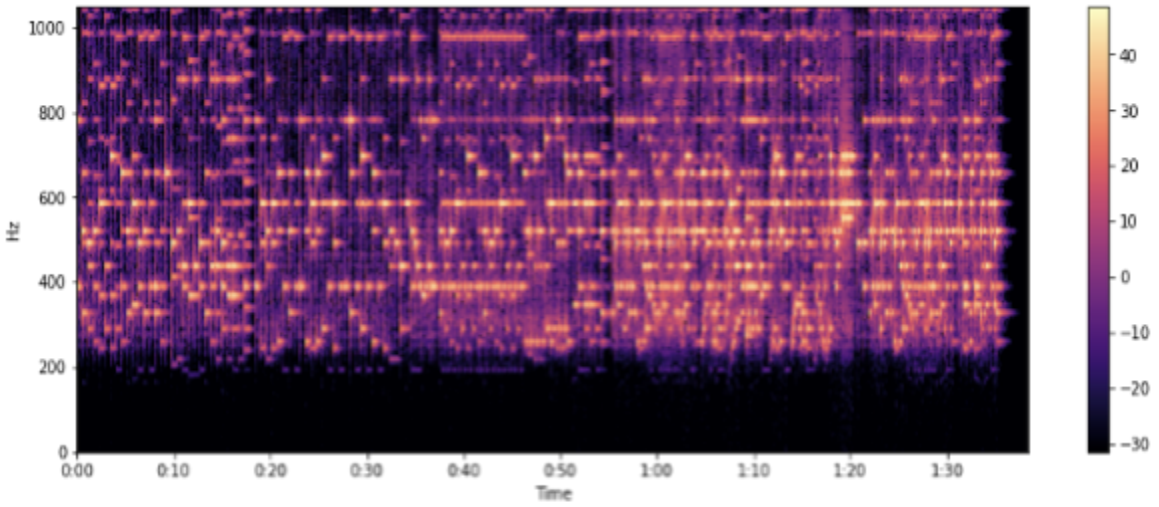
   For our 3 classification methods, we chose to use logistic regression, support vector machine (SVM), and a multi-layer perceptron neural network. For our SVM model, we chose an rbf kernel SVM due to the fact that our data was not directly linearly separable. For our final SVM model, we chose a hyperparameter model of gamma=10 and C=1. For our neural network model, we chose a 100-layer model using stochastic gradient descent (SGD) with a learning rate of 0.01 for a maximum of 500 iterations. *Remark: If the loss reaches a certain tolerance, the training will terminate before the max iteration is reached.*

4. **Results:**

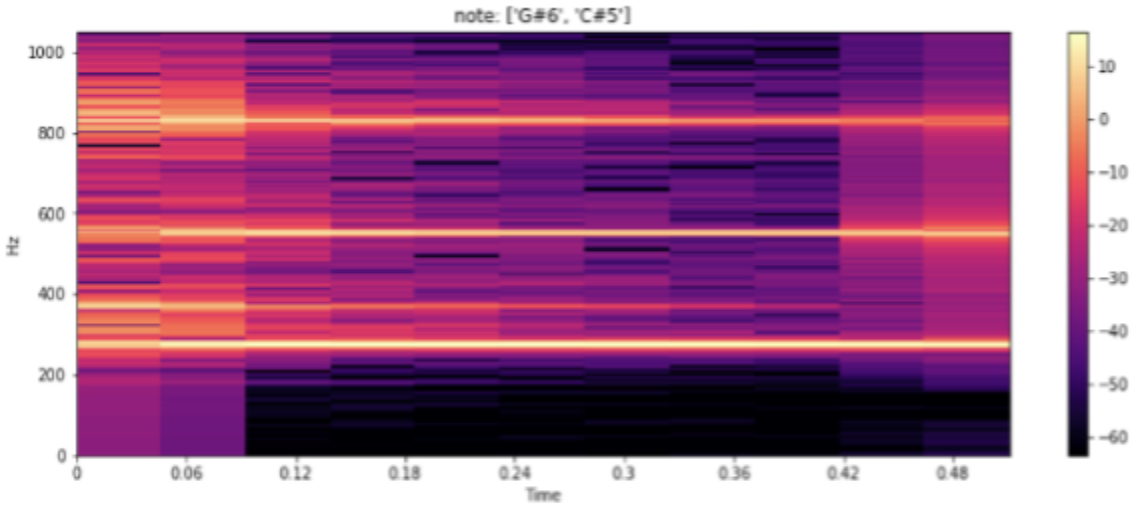
   Out of the three feature extraction methods we used, LDA performed the best. For LDA that reduced our data dimensionality down to 2 dimensions, we achieved an accuracy of 92.285%, 94. 252%, and 95.73% on the validation set on logistic regression, SVM, and neural network respectfully. On the test set, we achieved an accuracy of 90.898%, 91.095%, and 91.665%. To compare our results on another feature extraction method, we computed the classification accuracy using LSA that reduced our data dimensionality to 5 dimensions. On the validation set using LSA, we achieved an accuracy of 86.315%, 87.387%, and 87.387% on logistic regression, SVM, and neural network respectfully. On the test set, we achieved an accuracy of 86.196%, 89.595%, and 89.495%. As shown by our results, using LDA as a feature extraction method achieved much better results compared to LSA for predicting note values in piano rendition instrumentals.

Appendix

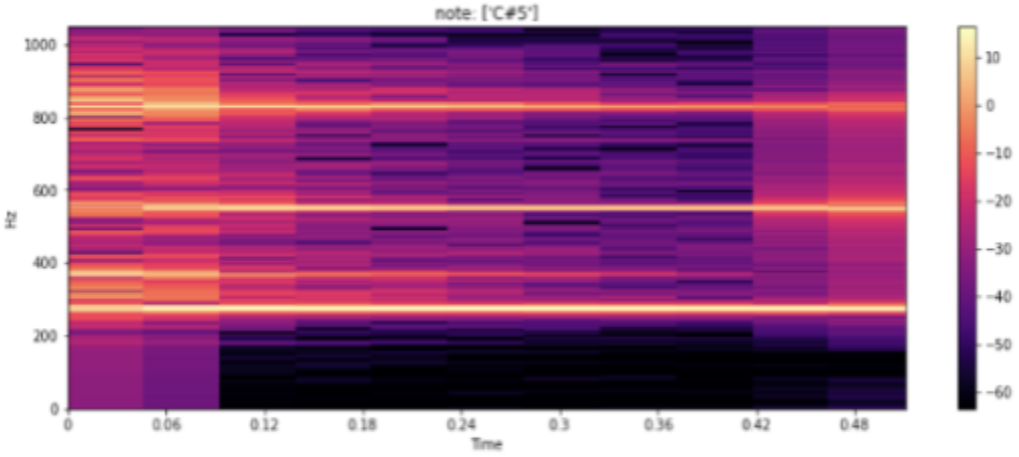
Figures



Spectrogram of a whole song being played

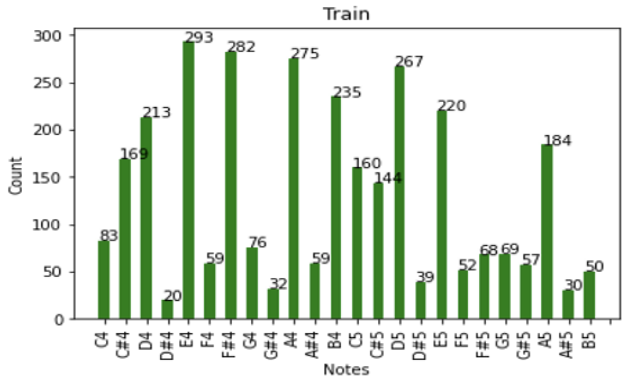


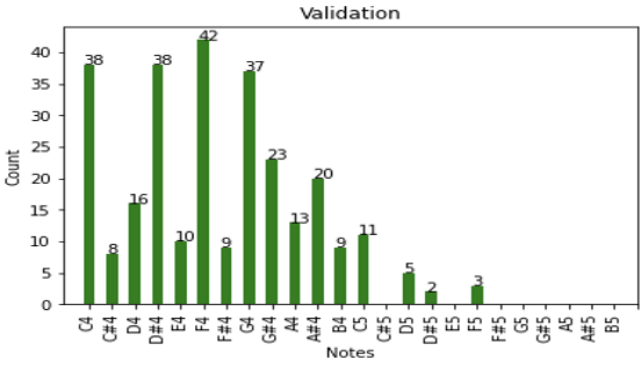
Spectrogram of Note G#6 and C#5 being played at the same time.

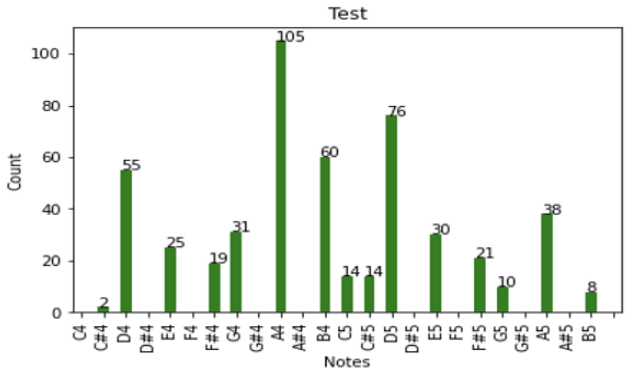


Spectrogram of note C#5 being played.

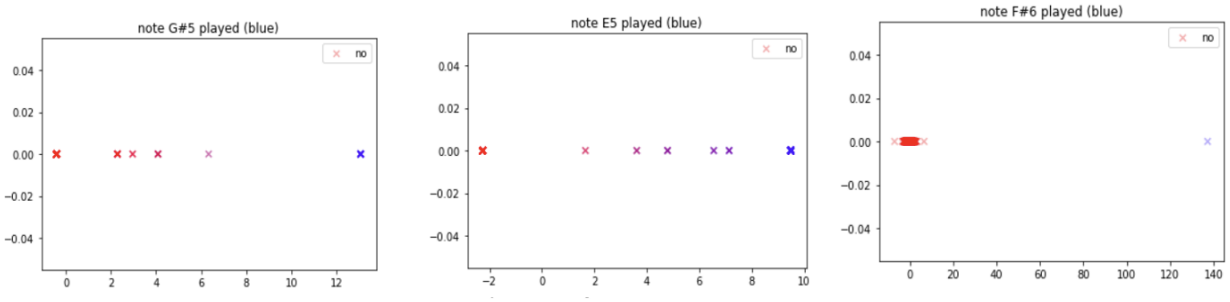
Note Distribution of train, validation, test sets:



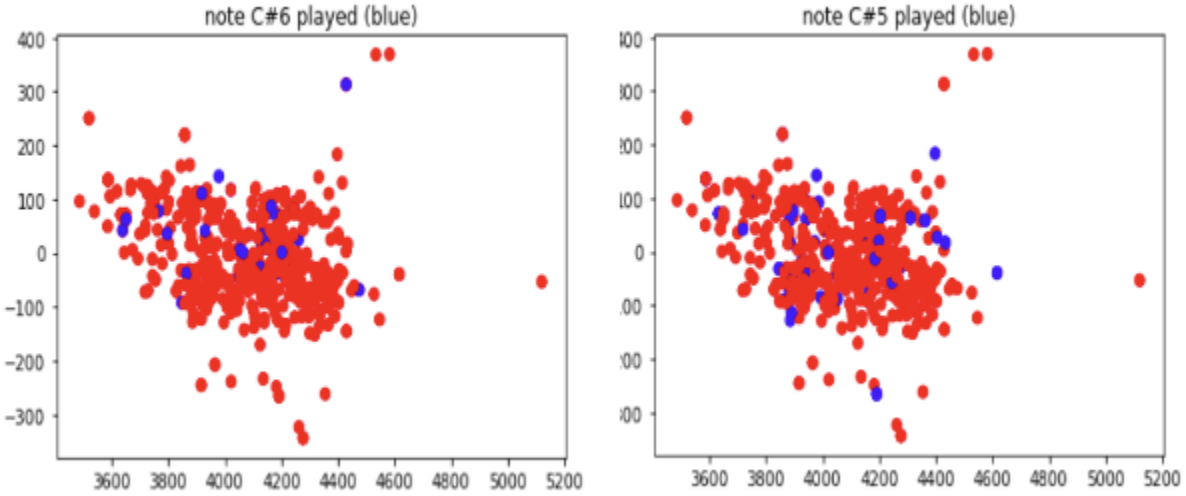




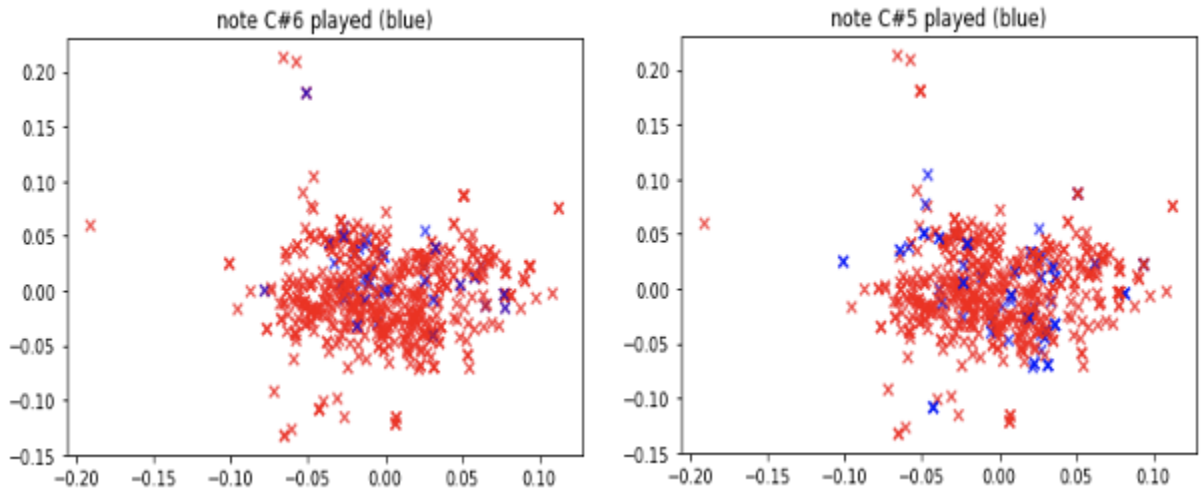
LDA Visualization:



LSA Visualization:



ICA Visualization:



Tables

Classification accuracy using LDA feature Extraction:

	Logistic Regression	SVM	Neural Network
Validation Accuracy	93.285%	94.252%	95.732%
Test Accuracy	90.898%	91.095%	91.665%

Classification accuracy using LSA feature Extraction:

	Logistic Regression	SVM	Neural Network
Validation Accuracy	86.315%	87.387%	87.387%
Test Accuracy	86.196%	89.595%	89.495%

Code

Please visit https://github.com/leronjulian/18752_project to view relevant software code for our project. Software code is also attached in the zip file.

References

[1] https://thehighestproducers.com/midi-file/ultimate-list/