A Temporal Approach to Unsupervised Anomaly Detection

Ashlae Blum

POSTER-21-053-STUDENT

Motivation

The problem of unsupervised anomaly detection of audio events remains a complex endeavor in the global AI community. To-date, much research has focused on frequency analysis to compare normal and anomalous audio data. However, we believe there is unexplored terrain with respect to temporal analysis, also known as onset detection. We engage in an experimental approach to understanding the role of percussive density in determining the sound quality of audio.

Approach and Objective

Using low-level techniques informed by a familiarity with music information retrieval, we determine the quality of audio features present in the dataset using the Librosa package for audio analysis. We then train an autoencoder to determine the anomaly scores of the data based upon frequency information. A dataset of temporal information is constructed and a LSTM is used to determine anomaly scores.

Spectral Features of Audio

The frequency domain of audio signals can be analyzed through a variety of methods. In this study we consider the Short Time Fourier Transform and Mel Frequency Cepstral Coefficients as a measure of power distribution across frequency bands.

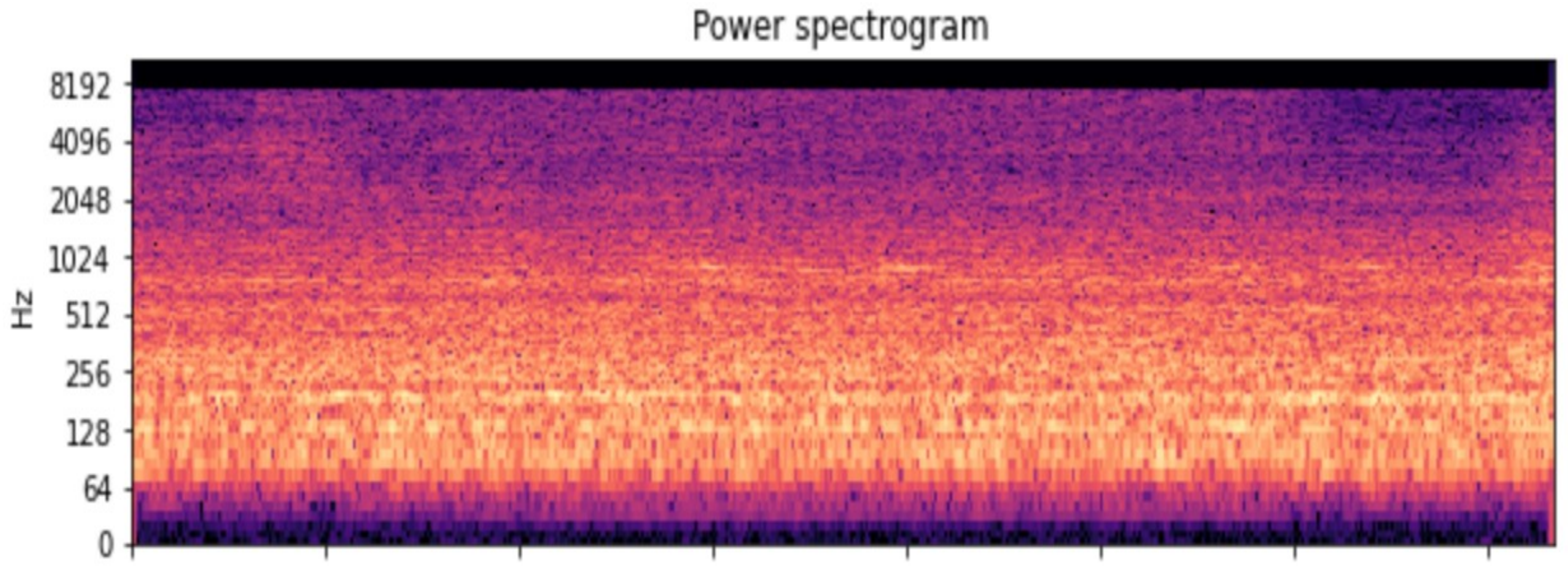

Fig 1: STFT power spectrogram of 10s anomalous audio file

Temporal Features of Audio

Features of interest we consider for temporal analysis are Uniform Tempo and Predominant Local Pulse. These are used to describe the occurrence of audio events, also known as Onset Detection.

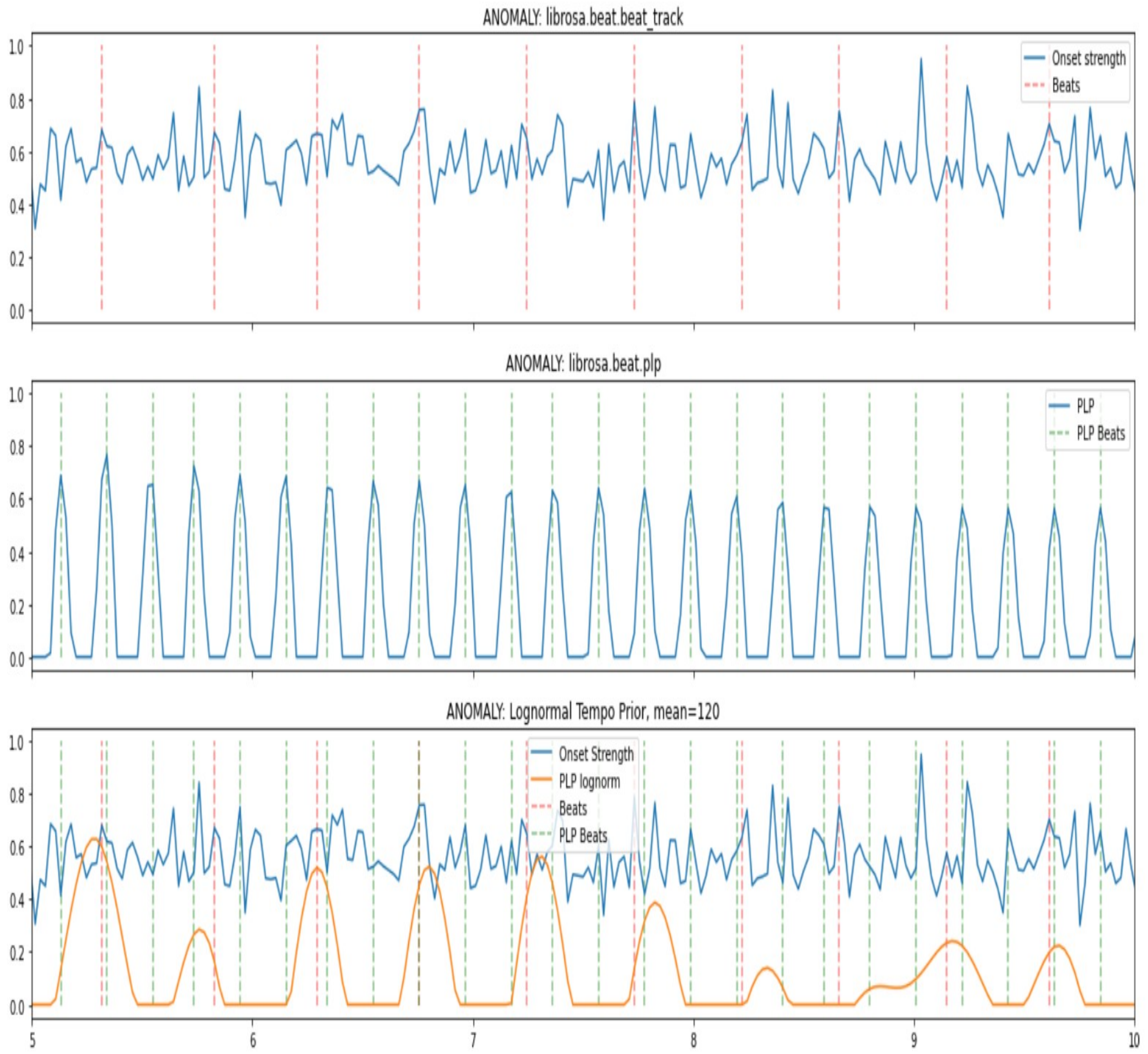

Fig 2: Onset strength, tempo, and PLP of anomalous audio

Methods

Datasets are constructed for the spectral and temporal data using Librosa and Pandas. An autoencoder is used to analyze the MFCCs. Using a sliding window of 128 mel bins and a hop length of 512, the 9-layer neural network compares spectral data to determine self-similarity over 100 epochs of training. The model produces an anomaly score for each audio file that represents the degree to which the network has determined the file is anomalous. Area Under Curve is then computed as a measure of model accuracy.

Results and Conclusions

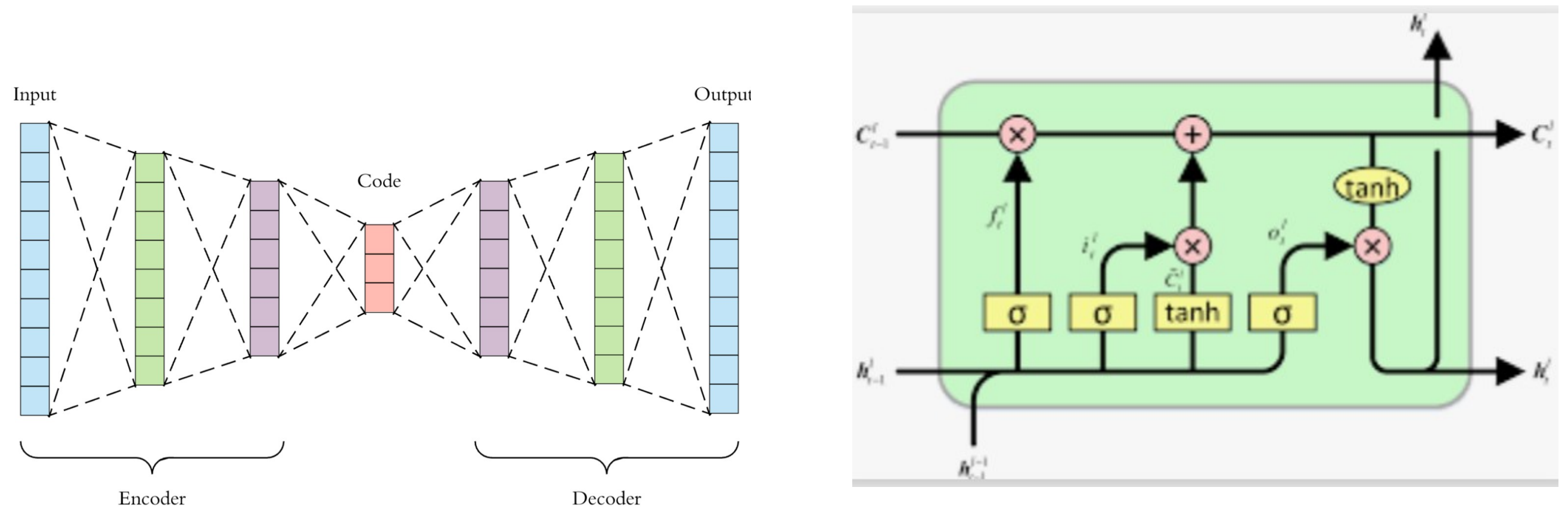

Fig 3: Autoencoder & LSTM neural networks

Using the autoencoder to analyze the MFCCs of the audio, we compute an AUC of 0.7779 for the model. This is a measure of the accuracy with which the neural network determines whether or not an audio file contains anomalies. Through experimental methods, we define an approach to analyzing temporal information by computing the average Tempo and Predominant Local Pulse of each audio file. This information is used to construct a dataset in a similar fashion to the MFCC datasets. However, the size of the temporal data is in fact non-uniform. That is, different audio files have different numbers of datapoints, resulting in different lengths of values within the dataset. As such, we must construct a different type of model to receive the data. We determine that a Long Short Term Memory network will be an appropriate model to train the data, since it is able to accept variable sized data.

Proposal for Future Research

We would like to propose a continuation of this research in the development of an LSTM to train the temporal audio data. There are several other extensible audio properties that could be explored and correlated. Of particular interest would be to compare spectral and temporal results.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.