Audio Classification of Musical Instruments with a Multimodal Neural Network

By:
Landon Buell[1]

Advised By:
Dr. Kevin Short[2] **, Dr. Maurik Holtrop**[1]

A thesis presented for the degree of
Bachelor of Science in Physics

December 2020

[1]Department of Physics and Astronomy
[2]Department of Mathematics and Statistics

University of New Hampshire
Durham New Hampshire, USA

Abstract

The task of recognizing the source of a sound wave in a digital audio file is trivial and almost effortless to a human being in many cases, but is far more difficult for a computer to do. In order to automate source source identification, we have designed and constructed a multimodal neural network that can classify audio files by what musical instrument the sound most resembles. The network uses two different input branches, one being a convolutional neural network (CNN), and the other, a multilayer perceptron (MLP). The CNN branch accepts input in the form of a two-dimensional spectrogram matrix, which encodes the energy distribution of an audio signal over frequency and time, and the MLP branch accepts input as a one-dimensional array of features derived from the physical nature of the sound wave. The spectrogram is produced by using a concatenation of multiple short-time Fourier transformations (STFT) of overlapping windowed time-space analysis frames. The feature vector implements techniques from digital signal processing to combine more explicit properties of the time-space and frequency-space information. To classify the audio files based on predicted sources, we aggregate and transform the activations from the final layers of each branch, and produce a single prediction label. We find that this architecture allows for reasonable classification performance, and demonstrates superior classification performance on average, when compared to either of it's constituent unimodal architectures. Training data for this experiment is from studio recordings of the Philharmonia Symphony Orchestra and the University of Iowa's Electronic Music Studios.

All source code and documentation for this project can be found at
https://github.com/landonbuell/Buell-Senior-Thesis

Contents

1 Introduction **3**
1.1 Introduction . . . 3
1.2 Methodology . . . 5

2 The Neural Network **7**
2.1 An Introduction to Neural Networks . . . 7
2.2 The Structure of a Neural Network . . . 8
2.3 Layers Used in this Classification Model . . . 9
2.4 Activation Functions Used in Network Layers . . . 16
2.5 Training The Model . . . 17
2.6 Multimodal Architecture . . . 26

3 Properties of Musical Instruments **31**
3.1 Idiophones . . . 32
3.2 Membranophones . . . 34
3.3 Chordophones . . . 34
3.4 Aerophones . . . 36
3.5 Other Generated Sounds . . . 39

4 Feature Selections **42**
4.1 Feature Space . . . 42
4.2 Spectrogram Features . . . 48
4.3 Time-Space Features . . . 53
4.4 Frequency-Space Features . . . 60

5 Evaluating Model Performance **65**
5.1 K-Folds Cross Validation . . . 65
5.2 Performance Metrics . . . 66
5.3 Tracking Metrics over a Period of Training . . . 70

6 Experimental Results **74**
6.1 Executing Cross Validation . . . 74
6.2 Comparing Results between Architectures . . . 75
6.3 Comparing Classification Scores within Each Class . . . 77
6.4 Discussion of Results . . . 83

7 Conclusion **86**

8 Acknowledgments **87**

1 Introduction

1.1 Introduction

The ability to classify time-series data into unique categories has become topic of much attention in the last few decades [9, 16]. From climate measurements, to financial reports, to audio-visual media, or to medicine, time-domain signals surround the modern world at every corner [32]. In each context, the general idea is to place information into distinct categories called *classes* based on inherent properties of the data. For example, doctors may wish to classify a patient as having a heart condition based on properties of their heartbeat or banks may wish to classify a purchase as fraudulent based on a history of spending habits. Often, the volume of information or complexity of the task makes it unreasonable for a human to do without the aide of some form of automation.

It has long been a goal of computer scientists and engineers alike to develop a program, algorithm, or machine that could perform a certain task to the level of human-being [1, 2, 13]. While humans are intricate organisms capable of recognition or classification in many contexts, we do not have the ability to process millions of raw numerical values in the course of a few seconds. Similarly, modern computers are also complex systems that can filter through exabytes of information on a daily basis, but are not generally known for being able to distinguish cats from dogs [2]. The compromise between humans and machines an algorithm called a *neural network* that combines the computation speed of a computer with the decision-making power modeled from human brain [4, 10, 13].

In this work, we have developed an automated neural network program that can classify digital sound waves as one of 37 musical instruments that it most closely resembles. The general use of machine learning algorithms for audio information classification or identification is well explored, and a topic of much modern research [9, 12, 16, 26, 35]. From previous work, it has been found that various types of audio classification are possible, and are still being expanded upon. Many papers detail possible sets of inputs for classifying speech against music or inputs for organizing audio-visual media based on content. In each case these inputs, called features, must often be tailored to meet the specific task at hand [32, 12].

It has been experimentally determined that the choice of features is critically important to classification success [9, 16, 26]. In our case, the nature of these features is motivated by the physics and mechanical properties of each musical instrument class. By exploring the physics of the vibrating parts of each instrument according to the Hornbostel-Sachs classification system, we gain a more thorough understanding of each instrument group. This allows us to develop properties within each class that are consistent and have low variability, and what properties differ between classes and show a higher variability. These differences are key to providing a set of features which allow a neural network to differentiate between classes of musical instruments [12, 35]. For our classifier model, we have developed 24 features from the time-series and frequency-series representation of the audio file data. These quantities

are assembled into a single axis array for processing. Additionally, we construct an image-like representation of the waveform and the 2D array is also used for proceessing. Both of these input modes represent a set of features that convey the nature of the waveform in two different modalities.

As the name implies, a neural network is a computation a model or graph that is loosely structured off of the biological brain [2, 13]. Where a biological brain uses physical connects through axons to exchanges chemicals to communicate information, a neural network uses a chain of mathematical operations to communicate information [4]. The earliest work on attempting to model cognitive processes goes back to 1943, where Warren McCulloch and Walter Pitts began to develop a model of the human brain using principles of mathematics [15, 13]. Nearly ten years later, a simple neural networks called the *perceptron* was developed by Frank Rosenblatt, which modeled connections of neurons through principles of linear algebra [2, 22]. Another decade after that, inspiration from the visual cortex gave rise to the convolution operation to be applied to digital image processing.

The design of our neural network function combines two distinct architecture types by using two distinct entry arrays which are processed, and a single output is formed. This type of neural network is referred to as a *multimodal* neural network [11, 18]. Our model contains a convolutional neural network (CNN) that processes an image-like representation of an audio waveform, and a multilayer perceptron (MLP) that processes a list-like representation of an audio wave form. Each arm of the network is it's own computational graph that is designed to process it's respective inputs types. The activations at the end of each layer are concatenated into a new layer, which is further processed and then used to generate a single prediction. This output that is aggregated from two inputs shows on average, superior classification performance when compared to either of it's constituent unimodal architectures alone.

1.2 Methodology

This project includes a number of steps from taking raw input data to the final classification program. In this section, we provide a very high level outline of the parts of this projects and show how each one influence the final product. Additionally, we indicate the sections in this thesis where the particular material is introduced and developed much more thoroughly.

1.2.1 Designing the Function

To automate the process of classifying audio files, we require a program or algorithm that can make practical decisions based on a prescribed set of information. This is to say that we need a *function* to process an input or set of inputs, and return a predicted label that corresponds to a musical instrument. Consider the biological process of hearing a sound wave and matching it to a source. We can model this behavior by some unknown function F and approximate it with the function F^*. This approximate function is composed of smaller functions that collectively work together to direct the flow of information and allow for the mapping of the contents of a sound file to a potential source as to mimic the behavior of F. For a set of inputs $\vec{x} = \left\{x_0, x_1, x_2, ..., x_{p-1}\right\}$ and a set of classes 0 through $k-1$, we denote this function as:

$$F^* : \vec{x} \rightarrow \left\{0, 1, 2, ..., k-2, k-1\right\} \tag{1}$$

The smaller functions that comprise F^* are called *layers* [13, 4]. Each type of layer performs a different operation and processes different types of input. In secs. (2.1) and (2.2), we introduce the nature and structure of a neural network and how it can use used in part to solve this problem. We discuss it is inspired from a biological brain and show how different layer types are influenced from the brain's functions as well as how they are used in this problem. Using information about the problem, we carefully design the inputs, outputs and structure of the neural network function to be able to complete this task in a practical manner.

1.2.2 Collecting and Pre-processing Raw Data

In order to train the neural network to be able to identify musical instrument sources, we need a suitable data set to present to the model as training and validation data. The models uses this data and corresponding labels in the training process to modify it's function to allow for it to complete the task. University of Iowa Electronic Music Studio, and the London-Based Philharmonia Orchestra each have a large collection of publicly available audio files [23, 33]. These files contain short segments of musical instruments performing a single note, and are labeled according to the instrument that produced the sound. The files and labels make up the core of the data set that we use to train the model.

To ensure that these data sets are formatted consistently, we have designed a MATLAB program to read each audio file from it's original format, *.AIF*, *.MP3*, or similar, and rewrite the data as a new *.WAV* files, sampled at 44.1 kHz with a 24-bit depth. This ensures that

all data will have a consistent format when features are extracted. This stage is detailed in section (4.1.3). After the samples are formatted, we have produced a Python script with iterates through each new file and assigns a class label based on the title of the instrument. This has been design to account for the specific naming cases that arise. We assemble a dictionary-like structure of all training file with organizes the local file location of each sample and provides an integer and string label for all samples.

1.2.3 Designing Classification Features

The performance of a neural network is largely dependent of designing an appropriate set of predictors or features which represent the raw input given to the model. These inputs are designed to represent prolific properties of wave forms that can be expressed by a single or multiple numerical values. We detail these features in section (4). We use tools from physics, mathematics, and signal processing to define and explore a comprehensive set of features that enables a high performance of the classifier.

1.2.4 Designing A Complementary Network Architecture

With the appropriate set of features designed, and the number of output categories determined, we can organize the structure of the neural network function. The shape of the network defines the flow of information and the hypothesis space. The nature of all possible solutions that the trained model could arrive at is subject to the architecture [4]. We construct a multimodal neural network, explored in section (2.6), that processes two distinct entry points derived from the same audio sample. This network is designed to evaluate each input independently, and concatenate the results to produce a single outputted prediction. This hybrid architecture combines the predictive power of the both a multilayer perceptron and a convolutional neural network. Our neural network was implemented using the tools available with the *Tensorflow* module [31].

1.2.5 Testing and Evaluating Network Performance

We execute a K-Folds cross validation algorithm to evaluate the performance of the neural network across overlapping subsets of data. This ensures that the model can perform consistently and reliably given small changes in initial conditions. In this stage, we also examine the value of hyper-parameters, activation functions, and layer widths to best compliment the properties of the data set. This process is repeated and expanded upon until we have produced a model with a sufficient performance. We explore this in section (5). Results of all metrics are computed with the *numpy*,*scipy*, and *scikit learn* python modules [**?**, **?**, **?**]. All performance visualizations are produced with the *matplotlib* python package [**?**].

2 The Neural Network

2.1 An Introduction to Neural Networks

Because of the complexity of sound recognition, the challenge arises as to how to build some sort of program that could function at a level above general procedural or explicit instructional rules. Rather than *hard-coding* a set of conditions or parameters for a classification program, we seek an solution that allows for a computer to *learn* from labeled training examples in order *teach* itself [1, 17]. We choose to employ a neural network, which a computer algorithm that is designed to update itself to perform a task with increasing proficiency as it is presented with new information [3, 4, 10, 17].

As the name implies, early neural networks were inspired from from the human brain. Former YouTube video classification team lead, and current machine learning consultant, Aurelien Geron writes about the relationship between biological brains and mathematical neural networks [2]:

> Birds inspired us to fly, burdock plants inspired velcro, and nature has inspired many other inventions. It seems only logical, then, to look to the brain's architecture for inspiration on how to build an intelligent machine.

The result is a computer program that is reminiscent of the a brain in that it structured much like the brain, and can*learn* from new data. We model operations with collections are artificial neurons that are capable of connect and interacting with other neurons [10]. Where a biological brain uses connections of chemical and electrical exchanges to process input and make decisions, the computational counterpart uses arithmetic and numerical exchanges to similarly process input and make decisions [10, 1]. Where a brain uses physical connection of neurons through axons, the neural network uses mathematical connections of neurons through mathematical operations [2, 10]. We call the numerical value contained within an artificial neuron the *activation* of that neuron [4, 2].

When a listener hears a sound, millions of neurons in their brain exchanges those chemical and electrical signals in such a way as to be able to identify the source of that sound [2, 32]. More conceptually, we can say that a listener is presented with a set of information x, which is air molecules vibrating near their tympanic membrane [34, 20]. That information is then passed into the brain where some function F is used to model the incredibly complex set of neural connections. The output of that function is some label y, which the listener associates with a particular source [20]. Similarly, we construct a neural network to accept a set of predictors x, which is passed into a function F^* that is use to approximate F, and produce an output y^*, which predicts the source of the audio sample.

The neural network function F^* uses a set of parameters Θ, to map those inputs, x to outputs, y [4, 8, 32]. It is the process of updating the elements in the object Θ which allow us to train a neural network to perform the desired task. Similarly, it is the process of

organizing and connecting neurons in the brain that allow humans to perform any specific task [2, 10]. Over the course of their lives, humans will be able to learn to match sounds to sources with relative ease, given the appropriate labels. In a similar manner, our neural network is presented a collection of predictors from sound files and a complementary set of labels indicating the source of each sound. This type of task is called *classification*, which involves mapping the input data, x to one of k possible *classes*, each represented by an integer, 0 through $k-1$. Each of the k classes represents a particular musical instrument or sound source that could have produced the sound-wave in the audio file [4, 13, 32]. In this section, we explore the motivation, structure, and training process of a neural network. We implement our neural network with the *Tensorflow* Python package [31].

2.2 The Structure of a Neural Network

A Neural Network is a model of a mathematical function, composed of several smaller mathematical functions called *layers* [4, 13]. Each layer represents an operation that takes an input, and returns an output . The exact nature of this operation can be very different depending on the layer type or where it sits within the network. Typically, a neural network is characterized by a chain-like structure where the output of one layer is fed directly into the input of the next layer [2, 4]. It is this process of transforming inputs successively in a particular order until an output is attained that makes up the core functionality of a neural network [2, 13]. The output of the final layer encodes the "decision" of the model given a particular input.

Other than inspiration from the brain, Neural Networks are named *networks* due to their successive chained or nested composition nature. A standard representation of a neural network is often a linked list or computational graph structure [4]. This maps out exactly how the repeated compositions are structured and organizes the flow of information in more complicated networks. Each node in the computational graph represents a *layer*, which executes a particular prescribed mathematical function [4]. In the case of a *linear, feed-forward* network, such as the one that we implement, information passes in a single direction to produce an output. We can represent this as a graph:

$$F^*(x) = x \rightarrow f^{(0)} \rightarrow f^{(1)} \rightarrow ..... \rightarrow f^{(L-2)} \rightarrow f^{(L-1)} \rightarrow y^* \tag{2}$$

or as a nested function:

$$F^*(x) = f^{(L-1)}\big[f^{(L-2)}\big[...f^{(1)}\big[f^{(0)}\big[x\big]\big]\big]...\big] = y^* \tag{3}$$

The number of layers in a neural network is referred to as the network *depth*. The dimensionality of each layer is referred to as the network *width* [2, 13]. A network model that contains L unique layers is said to be an L-Layer Neural Network, with each layer indexed by a superscipt, (0) through $(L-1)$. Layer (0) is said to be the *input layer* and layer $(L-1)$ is said to be the *output layer* [2, 13].

Due to the chained nature of the model, the activations from some layer, $(l-1)$, are used to directly produce the activations for the next successive layer (l). The function that represents a layer (l) is given by:

$$f^{(l)} : x^{(l-1)} \in \mathbb{R} \rightarrow x^{(l)} \in \mathbb{R} \tag{4}$$

The array of activations, $x^{(0)}$, is the raw input given the neural network, called *input* activations. Conversely, $x^{(L-1)}$ is referred to as *output* activations [2, 8, 13].

Below we outline a standard "forward pass" algorithm for a general feed-forward deep neural network. We presume each layer to be a node in a computational graph that contains a method "*call*" which executes the layer's main function as in Eq. (4). Additionally, each node contains a pointer "*next*" which gives the next layer in the network chain. Each layer should also store it's linear activations, $z^{(l)}$ and it's output activations $x^{(l)}$ for training purposes. This algorithm is adapted from Goodfellow, [4].

Algorithm 1 Forward propagation system in a standard deep neural network. Each layer is presumed to be a node in a linked computational graph. This example has been setup to assume one input layer, and one output layer. Practical implementations should include mini-batches of data as opposed to a single sample.

Require: Set of Layer functions $f^{(i)}$, $i \in [0, 1, 2, 3, ..., L-1]$.
Require: Input sample - $x^{(0)}$
  Set the current layer to be the input layer:
  $f^{(i)} \leftarrow f^{(0)}$
  for Layer $i \in [1, 2, 3, ..., L-1]$ **do**
    Run "Call" Method with current activations, transform activations:
    $x^{(i)} = f^{(i)}.\text{call}(x^{(i-1)})$
    Move to the next layer in the network:
    $f^{(i)} \leftarrow f^{(i)}.\text{next}()$
  end for
  Return the network's output, which are the output activation of the final layer
  return $x^{(L-1)}$

2.3 Layers Used in this Classification Model

As stated previously, a neural network is composed of a series of functions that are called in succession to transform features (an input) into predictions (an output). As shown in Eq. (4), each function feeds directly into the next as to form a chain-like computational graph [4]. There are multiple type of layers that we use to construct the network in this project, but the two most important are called a *dense* layer and a *convoltuion* layer. These layers each contain a set of parameters which makeupa subset of the trainable elements in Θ. These two layer functions can be divided into two portions: (i.) a Linear transformation, with a bias addition, and (ii.) an element-wise activation transformation. The Linear transformation

can be given by Eq.(5) in the case of a 1D Dense layer (2.3.1) or Eq. (6) in the case of a 2D Convolution layer.

$$z^{(l)} = W^{(l)}x^{(l-1)} + b^{(l)} \tag{5}$$

$$z^{(l)} = \left(W^{(l)} * x^{(l-1)}\right) + b^{(l)} \tag{6}$$

Where $W^{(l)}x^{(l-1)}$ denotes a matrix-vector product, and $\left(W^{(l)} * x^{(l-1)}\right)$ denotes a convolution product. In both cases, $W^{(l)}$ is a *weighting-matrix* or *weighting-kernel* for layer (l), $b^{(l)}$ is the *bias-vector* for layer (l), $z^{(l)}$ are the *linear-activations* for layer (l) and $x^{(l-1)}$ is the final activations for layer $(l-1)$.

Step (ii.) is usually given by some *activation function*:

$$x^{(l)} = \sigma^{(l)}\left[z^{(l)}\right] \tag{7}$$

Where $x^{(l)}$ is the final activations for layer l and $z^{(l)}$ is given in equation (5). $\sigma^{(l)}$ is some activation function which introduces the ability to account for complicated non-linear decision boundaries which the neural network uses to make predictions. The combination of Eq. (5) and Eq. (7), for a layer l , make up the function represented by that layer, as in Eq. (4). In this section, we detail the types of layer functions that are used to produce the classification model in this project.

2.3.1 Dense Layer

The Linear Dense Layer, often just called a *Dense* Layer, or *Fully-Connected* layer, was one of the earliest and most common function types used in artificial neural network models [3, 13, 15]. A dense layer is composed of a layer of *artificial neurons*, each of which holds a numerical value within it, (usually a double-precision floating point number) called the *activation* of that neuron. The activation of a dense layer are given by a linear combination of the layers inputs. This idea was developed from McCulloch and Pitts' work [15] in attempting to model cognitive processes as mathematical functions by connecting neurons through a series of weighting values. The idea was expanded upon by Frank Rosenblatt in 1957 by combining many dense layers together to form a multilayer perceptron (MLP), which will make up a portion of the network that we have implemented. [2, 13, 10].

We model a layer of neurons as an array or vector of floating-point numbers. Suppose a layer (l) contains n artificial neurons- we denote the n-dimensional array that holds those activations as $x^{(l)}$ and is given by:

$$\vec{x}^{(l)} = \Big[x_0, x_1, x_2, \ldots, x_{n-2}, x_{n-1}\Big]^T \tag{8}$$

The activation of each entry is given by a linear-combination of activations from the previous layer, Eq.(5) and the transformation in Eq.(7).

Suppose the layer $(l-1)$ contains m neurons. Then the weighting-matrix, $W^{(l)}$ has shape $m \times n$, the bias-vector $b^{(l)}$ has shape $n \times 1$. We can denote the function in a similar manner to Eq.(4) as:

$$f^{(l)} : x^{(l-1)} \in \mathbb{R}^{(m \times 1)} \rightarrow x^{(l)} \in \mathbb{R}^{(n \times 1)} \tag{9}$$

Thus for a dense layer l, the exact values of each activation is given by [2, 13]:

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix}^{(l)} = \sigma^{(l)} \left(\begin{bmatrix} w_{0,0} & w_{0,1} & \dots & w_{0,m-1} \\ w_{1,0} & w_{1,1} & \dots & w_{1,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n-1,0} & w_{n-1,1} & \dots & w_{n-1,m-1} \end{bmatrix}^{(l)} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{m-1} \end{bmatrix}^{(l-1)} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix}^{(l)} \right) \tag{10}$$

or more compactly:

$$x^{(l)} = \sigma^{(l)} \Big(W^{(l)} x^{(l-1)} + b^{(l)} \Big) \tag{11}$$

Eq. (11) is generally referred to as the *dense layer feed-forward equation* [4]. Below, we detail pseudo-code for what a "Call" method for a dense layer node in a computational graph may look like [2, 31] Compare this process with Eq. (11).

Algorithm 2 Typical "Call" method for a dense layer in a neural network that contains n neurons/nodes. This example shows the computation over a single input $x^{(l-1)}$ but a practical implementation should include mini-batches of samples.

Require: $x^{(l-1)} \leftarrow$ Input activations shaped into $(m \times 1)$ column vector
Require: $n \leftarrow$ Number of nodes in this dense layer
Require: $\sigma^{(l)} \leftarrow$ Activation function for this layer
Require: $W \leftarrow$ a $(n \times m)$ matrix of weights
Require: $b \leftarrow$ a $(n \times 1)$ vector of bias values
Evaluate linear activation values by computing the standard matrix-vector product of $W^{(l)}$ and $x^{(l-1)}$, and adding the bias neuron value:
$z[i] \leftarrow b[i] + \sum_{j=0}^{n-1} W[i,j] x^{(l-1)}[j]$
Evaluate output activation values by applying the activation function to each element in the linear activation column vector z:
$x^{(l)}[i] \leftarrow \sigma^{(l)}\big(z[i]\big)$
Store both activation arrays for back-propagation. Return the output activations:
return $x^{(l)}$

2.3.2 2-Dimensional Convolution Layer

Convolution layers emerged from studying the brain's visual cortex, and have been used from image recognition related tasks for around 40 years. [2, 13]. The layer function get's it's name from it's implementation of the mathematical *convolution* operation. The 2D discrete-convolution of function or 2D array $A[i,j]$ and $B[i,j]$ is given by [4]:

$$C[i,j] = (A * B)[i,j] = \sum_{u} \sum_{v} A[i,j] B[i-u, j-v] \tag{12}$$

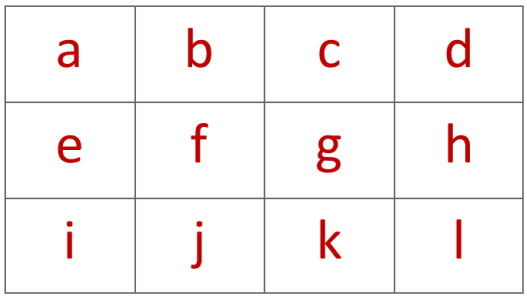
Note that convolution is commutative: $(A * B) = (B * A)$. We implement a convolutional layer $f^{(l)}$ by creating a series of K weighting matrices, called *filters*, each of size $m \times n$, where $m, n \in \mathbb{Z}$. Often we choose $m = n$ and we call the shape the *convolution kernel size* [13, 4]. The step size in each of the two dimensions is known as the *stride size*, which we choose to be 1 in both dimensions.

In the case of a 2D input array, $x^{(l-1)} \in \mathbb{R}^{(M \times N)}$, the convolutions filters "step" through the input data and repeatedly compute the element-wise product of each $m \times n$ weighting kernel, and the local activations of the $x^{(l-1)}$ array. Each of the K filters are used to generate a $m \times n$ feature map from the input activations. For an appropriately trained network, some filters may result in detecting vertical lines, horizontal lines, sharp edges, areas of mostly one color, etc. [2, 13].

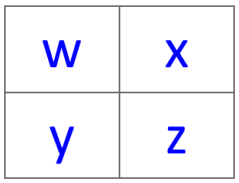
In general, for an $M \times N$ input, with kernel size $m \times n$, with stride size 1×1, and K feature maps.

$$f^{(l)} : x^{(l-1)} \in \mathbb{R}^{(M \times N)} \rightarrow x^{(l)} \in \mathbb{R}^{([M-m+1] \times [N-n+1] \times K)} \tag{13}$$

As an example, consider a single filter map over input $x \in \mathbb{R}^{(3 \times 4)}$ and filter map $W \in \mathbb{R}^{(2 \times 2)}$ as shown in Fig. (1):



(a) Input Activations, $x^{(l-1)}$



(b) Convolution Filter, $W_k^{(l)}$

aw + bx + ey + fz	bw + cx + fy + gz	cw + dx + gy + hz
ew + fx + iy + jz	fw + gx + jy + kz	gw + hx + ky + lz

(c) Convolution Result, $x_k^{(l)}$

Figure 1: The result of convolving an input (a) with an filter map (b) is a new set of activations (c). This Image was adapted from Goodfellow, pg. 325 [4]

Note that formal implementations include a bias vector and an activation function.

For 2D convolution layer (l), given input $x^{(l-1)}$, we compute the activations of the k-th filter $x_k^{(l)}$ using feature map $W_k^{(l)}$ and bias neuron $b_k^{(l)}$ as [4]:

$$x_k^{(l)}(i,j) = \sigma^{(l)}\left[b_k^{(l)} + \sum_{u=0}^{m-1}\sum_{v=0}^{n-1} W_k^{(l)}(i,j)x_k^{(l-1)}(i-u, j-v)\right] \tag{14}$$

This operation repeats for each of the K feature maps. Each maps has it's own $n \times m$ weighting matrix and appropriately shaped bias .

The convolution layers allows for several advantages over the dense layer. Among these are (i) *sparse-connectivity*: not every single activation (pixel) is connect to every single output pixel, so it is more computationally efficient, (ii) *positional invariance*: key features can be identified regardless of where they are in the layer, and (iii.) *Automatic feature detection* as the training process will update the filters to identify dominant features in the data without human instruction [2, 4, 13]. Below we detail what the "Call" method for a 2-dimensional convolutional neural network may look like. Compare this procedure with Eq. (14).

Algorithm 3 Typical "Call" method for a 2-Dimensional Convolutional layer in a neural network that uses K filters, of $m \times n$ kernels, with an assumed stride size of 1×1. This example shows the computation over a single input $x^{(l-1)}$ but a practical implementation should include mini-batches of samples.

Require: $x^{(l-1)} \leftarrow$ Input activations shaped into $(M \times N)$ matrix
Require: $K \leftarrow$ The number of filters (feature maps) to be used.
Require: $\sigma \leftarrow$ Activation function for this layer
Require: $W \leftarrow$ array of weight kernels containing $W_k, k \in [0, 1, ..., K-1]$ each shaped $m \times n$
Require: $b \leftarrow$ a $(1 \times K)$ vector of bias values, $b_k, k \in [0, ..., k-1]$
  for Each feature map, $k \in [0, 1, 2, ..., K-2, K-1]$ **do**
    Evaluate linear activation values by computing the 2D convolution-product of W_k and $x_k^{(l-1)}$:
    $z_k[i,j] \leftarrow b_k + \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} W_k[u,v] x_k^{(l-1)}[u-i, v-j]$
    Evaluate output activation values by applying the activation function to each element in the linear activation matrix z:
    $x_k^{(l)}[i,j] \leftarrow \sigma\big[z_k[i,j]\big]$
  end for
  Store both activation arrays for back-propagation. Return the output activations from each of the K filters:
  return $x^{(l)}$

2.3.3 2-Dimensional Maximum Pooling Layer

A Maximum Pooling layers returns the maximum neuron activation in a two-dimensional group of neurons. In the case of 2D Max Pooling, we choose a kernel size to be $m \times n$, similar to the convolution layer, and extract the maximum value in each window, while stepping along according to a chosen stride size [13, 4]. Consider an input like in Fig. (1a). Using the 2×2 kernel on the 3×4 input, each box would then contain the maximum value of the input activations. We detail this in Fig. (2):

a	b	c	d
e	f	g	h
i	j	k	l

(a) Input Activations, $x^{(l-1)}$

Max(a,b,e,f)	Max(b,c,f,g)	Max(c,d,g,h)
Max(e,f,i,j)	Max(f,g,j,k)	Max(g,h,k,l)

(b) Output Activations, $x^{(l)}$

Figure 2: The result of 2D maximum-pooling an input array. This image was adapted from Loy, pg. 126 [13]

Pooling layers, such as maximum pooling, average pooling, or similar are typically placed after a layer or a set of layers of convolution [13]. The purpose of this arrangement is to reduce the number of activations by selecting only the largest valus, thereby dropping the width of the model and ensuring only features with large activation values are preserved [2, 13, 4]. This layer "non-trainable" because it does not use any weights or transformations, only a fixed procedure that cannot be updated. Below, we detail what the "Call" method for a 2D maximum pooling layer might look like.

Algorithm 4 Typical "Call" method for a 2-Dimensional Maximum Pooling layer in a neural network. We assume 2D input, but a practical implementation may include the need to loop over a high dimensional structure. For simplicity, we assume a stride size of 1×1. This example shows the computation over a single input $x^{(l-1)}$ but a practical implementation should include mini-batches of samples.

Require: $x^{(l-1)} \leftarrow$ Input activations shaped into $(M \times N)$ matrix
Require: $m \times n \leftarrow$ Pool height and width
Require: $P \leftarrow$ A temporary pool array of shape $m \times n$ to hold local activations
  for $i \in [0, 1, ..., M-2, M-1$ **do**
    for $j \in [0, 1, ..., N-2, N-1$ **do**
      Collect local activations in P, zero pad if necessary
      $P = x^{(l-1)}[i : i+m, j : j+n]$
      Find maximum value in P, and add to output
      $x^{(l)}[i, j] = \max\big(P\big)$
    end for
  end for
  Store the activation array for back-propagation, and return it
  return $x^{(l)}$

2.3.4 1-Dimensional Flattening Layer

A flattening layer is used to compress an array with two or more dimensions down into a single dimension. For a flattening layer (l), multidimensional activations in layer $(l-1)$ are rearranged down into an array such that each sample contains only one axis. Note that a batch of samples will still be two-dimensional. This is not like projecting the data into a lower dimension, but rather is the reorganization of values into an array of a row of column array. We can use function notation to express this for a single sample as:

$$f^{(l)} : x^{(l-1)} \in \mathbb{R}^{(M \times N \times ....)} \rightarrow x^{(l)} \in \mathbb{R}^{(MN... \times 1)} \tag{15}$$

The numerical value of each activation is left unchanged. For a layer with activation shape$M \times N$, the resulting activations are reshaped into $MN \times 1$ as shown in Eq.(15).

Flattening Layer are most commonly used to prepare activations for entry into dense layer or series of dense layers. For example, an image may contain two or three dimension, and

convolution over it will return two or three dimensional activations per sample. Each sample must be compressed into one dimensions so that it can be processed by dense layers. This layer also is "non-trainable" because it does no use any weights or transformations, only a fixed procedure.

2.3.5 1-Dimensional Concatenation Layer

The 1-Dimensional Concatenation Layer, also called 1D-Concat layer takes separate arrays of activations and combines them into a single array, while preserving the dimensionality of the sample. Consider the layer activations $\vec{a}^{(l-1)}$ and $\vec{b}^{(l-1)}$ with shapes $1 \times \alpha$ and $1 \times \beta$ respectively. They are the outputs of two different layers:

$$\vec{a}^{(l-1)} = \big[a_0, a_1, ..., a_{\alpha-1}\big] \qquad \text{and} \qquad \vec{b}^{(l-1)} = \big[b_0, b_1, ..., b_{\beta-1}\big] \tag{16}$$

The result of the concatenation is a new 1D-array, $\vec{c}^{(l)}$ with size $1 \times \alpha + \beta$:

$$\vec{c}^{(l)} = \big[a_0, a_1, ..., a_{\alpha-1}, b_0, b_1, ..., b_{\beta-1}\big] \tag{17}$$

We can denote this for n arrays with function notation:

$$f^{(l)} : x_a^{(l-1)} \in \mathbb{R}^{(1\times\alpha)}, x_b^{(l-1)} \in \mathbb{R}^{(1\times\beta)}, ... \rightarrow x_z^{(l)} \in \mathbb{R}^{(1\times\alpha+\beta+...)} \tag{18}$$

The concatenation layer is used to combine activations from two different layers into a single new layer. In this case of the model used in the project, we combine the outputs that result from the convolution branch and the perceptron branch to produce an single aggregated output. This process is detailed further in section (2.6). This layer also "non-trainable" because it does no use any weights or transformations, only a fixed procedure.

2.4 Activation Functions Used in Network Layers

Activation functions are an important factor in the behavior of neural networks [2]. For the convolution or dense layer, activation functions make up the second transformation step as in Eq. (7) which allows for the neural network to produce non-linear decisions. In this section, we detail the activation functions used in this classification model.

2.4.1 Rectified Linear Unit

The Rectified Linear Unit (ReLU) activation function acts element-wise on the activations in a given layer. If the activation of a neurons is non-negative, the value is untouched, otherwise a 0 is returned. For an input activation array x, ReLU is defined by:

$$\text{ReLU}\big[x\big] = \max\big(0, x\big) = \begin{cases} x & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

We provide a visualization of the function in Fig. (3).

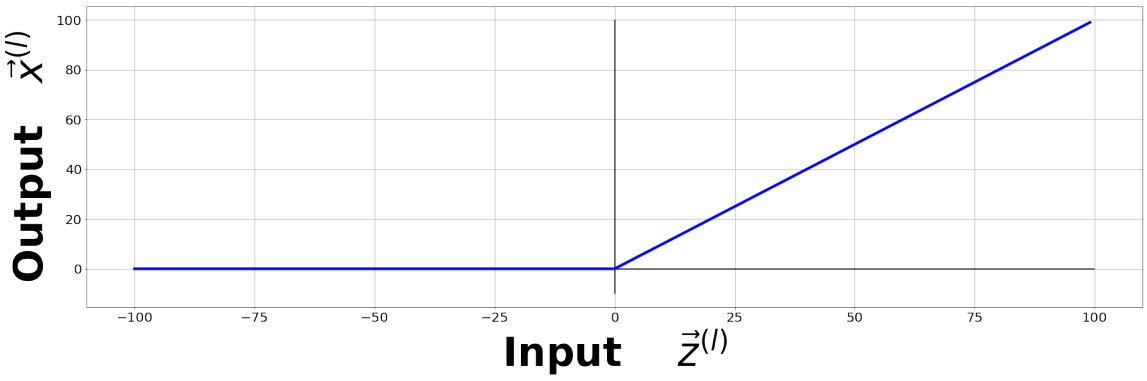


Figure 3: Rectifed Linear Unit (ReLU) activation function

For our chosen architecture, ReLU is applied to the activations in every single Convolution layer and Dense Layer, with the exception of the output dense layer.

2.4.2 Softmax

The softmax activation function is commonly used in the output layer of a multi-class classification network, such as the one we implement [2, 13]. When softmax acts on a vector of activations, the result a non-negative output vector with an L_1 norm of 1 [2, 4, 32]. The i-th element in a softmax activation function is given by:

$$\text{Softmax}\big[x\big]_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{20}$$

If Eq.(20) returns a vector of the form: $[0.75, 0.25]$, we interpret this as a sample having a 75% chance of belonging to class 0 and a 25% change of belonging to class 1. This tool is particularly useful is we are interested in determining the predictive confidence of a network. For example, if some if the two largest values in the output vector are very similar in magnitude, then the classifier is likely having a hard time differentiating between those two classes.

2.5 Training The Model

A neural network's purpose is to produce a function F^* that approximates an unknown function F, using a set of parameters, Θ. The model must have a procedure for updating the parameters in Θ to allow for a reasonably close approximation of F [4]. To better understand this, we turn to Tom Mitchell's explanation of a learning algorithm [17]:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if it's performance at tasks in T, as measured by P, improves with experience E.

Without any direct human intervention, a model must update itself to improve it's performance at a give task as new information is presented to it. To do this, the model must be constructed with a training procedure in mind.

We consider the set of parameters Θ as the set of values within each trainable layer's weighting matrix and bias vector such that:

$$\Theta = \left\{ W^{(0)}, b^{(0)}, W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, ..., W^{(L-2)}, b^{(L-2)}, W^{(L-1)}, b^{(L-1)} \right\} \tag{21}$$

The arrays $W^{(l)}$and bias vectors $b^{(l)}$ are not trainable parameters themselves, but the the floating-point entries within them are the values that can be adjusted. Each element in this set may contain hundreds or thousands of parameters, so we represent these indirectly as parameters within their respective layers. We choose this representation of Θ for notation simplicity. Note that not all layers have trainable parameters, and activations functions themselves are not trainable, but fixed functions. Modern neural networks can contain upwards of hundreds of thousands,or even millions of elements in Θ making a neural network a functions that exist is a very high dimensional parameter-space [2, 4, 10]. For the network that we have designed in this project, there are roughly $33,000$ parameters across 20 layers. In this sections we motivate and explore how the we can update the parameters in Θ as to train the neural network.

2.5.1 The Cost Function

Suppose we pass a training sample into the neural network. This sample is represented by the feature-vector $x^{(0)}$, with an expected outcome given by the one-hot-encoded vector y. After the network finishes processing, it's output activations, given by the vector $x^{(L-1)}$, also noted as y^*, represents the prediction for the class label. For a reasonably trained model, we expect the vector y^* to be "similar" to y, indicating that model has made a a reasonable prediction with the given input values. Alternatively, for an untrained network, y^* is not likely to be "similar" to y at all, indicating that the model has made a poor prediction with the given input.

To quantify the difference between the model's prediction, y^* and the expected output, y, we introduce a *cost function*, $J(y^*, y)$ [4, 8]. The cost function, also called a *loss* or *objective* function, compares y^* and y to return a single scalar value which measures the quality of the prediction. A high cost value indicates a *poor* prediction, and a low cost indicates a reasonable prediction. We can generalize this idea to consider that we want a trained neural network to produce a consistently low cost function value across samples in a data set.

The cost function itself, $J(y^*, y)$ can take many forms and is often dependent on particular task or data set provided [8]. For this k-classes classification task, we choose to use the *Categorical Crossentropy* (CXE) cost function. For a sample labeled by the one-hot-encoded

vector y and corresponding prediction vector y^*, the CXE objective value for a single sample is given by [1, 4, 32]:

$$\text{CXE}\big[y, y^*\big] = -\sum_{i=0}^{k-1} y_i \ln(y_i^*) \tag{22}$$

Thus, the average loss over a mini-batch of b samples is given:

$$\big\langle \text{CXE}\big[y, y^*\big] \big\rangle = -\frac{1}{b}\sum_{n=0}^{b-1}\sum_{i=0}^{k-1} y_i^{(n)} \ln(y_i^{*(n)}) \tag{23}$$

Suppose that a given sample belongs to class j in a k-classes classifier. Since the label vector, y is one-hot-encoded, all entries are zero except $y_j = 1$.

$$y = \big[y_0, y_1, ..., y_j, ..., y_{k-1}\big]^T = \big[0, 0, ..., 1, ..., 0\big]^T \tag{24}$$

Thus the sum in Eq. (22) contains mostly zero terms, with the only exception at index j, where we have $y_j \ln(y_j^*) = \ln(y_j^*)$. Since the vector y^* has been subject to the softmax activation function, we must have $y_j^* \in [0, 1]$. This means that taking the natural log of this value returns a negative number. Multiplying that by -1 returns a high loss when the activation of $y_j^* << 1$ and a low cost when $y_j^* \approx 1$. We visualize this relationship for a single sample in Fig. (4):

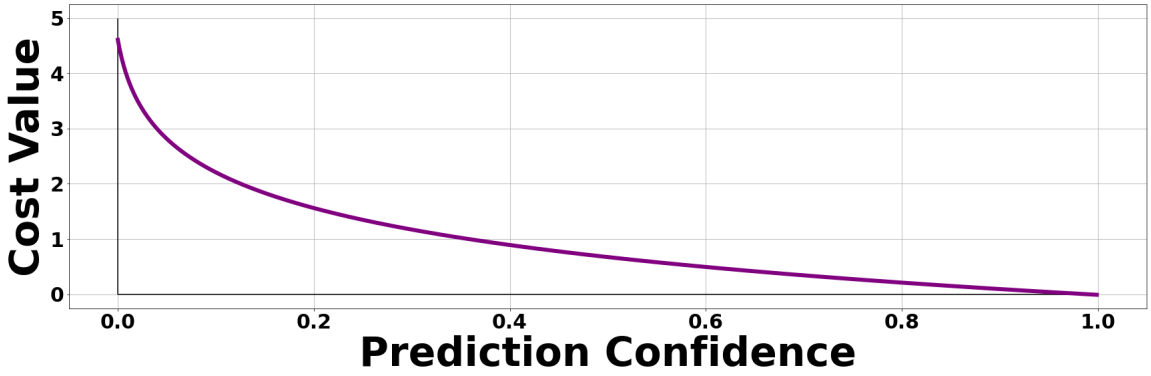


Figure 4: Plot of how $y_j^* \in [0, 1]$ affects the output values of the CXE cost function

By optimizing the parameters Θ in the network model, we allow for the output of a consistently low cost function to be produced across all samples in the data set. When a model produces consistently low cost values across new unseen samples, this means that $y^* \approx y$ and we consider this to be a *trained* or *fitted* neural network [4, 13, 17].

2.5.2 Gradient Based Learning

We have developed a method for quantifying the difference between a given output y^* and an expected output y with the inclusion of a cost function. The process of training the model

is to choose the parameters in Θ that allows for consistently low values of the cost function. We then treat the training process of training a neural network as a high dimensional *optimization* problem. A neural network uses *indirect optimization*, which contrasts from pure optimization. Deep Learning expert Ian Goodfellow describes the difference [4]:

> In most learning scenarios, we care about some performance measure P,... We reduce a different cost function, $J(\theta)$ in the hope that doing so will also improve P.

By choosing an appropriate cost function, and selecting the parameters Θ that minimize the average cost over a data set, we assert that $y^* \approx y$, but can only assume that doing so minimizes classification error and improves the performance metrics.

The cost value of a sample is dependent on the relationship between training labels y and the network output y^*. The output is given by the final layer activation $x^{(L-1)}$, which in turn are produced by the previous layer and so forth, as demonstrated in Eq. (4), This recursive nature combined with the dimensionality of the parameter object Θ makes an analytical solution to the optimization impractical [2, 4, 8]. We instead optimize the cost function with a numerical iterative method called *gradient descent* [13].

We can reduce the cost given a set of parameters, $J(\Theta)$, by repeatedly stepping each element in Θ. We step each element acording to the direction and manitude of each element in a gradient vector, $\nabla_\Theta J$, which is the gradient of the cost function with respect to each parameter in Θ. We express the gradient of J with respect to the parameters in Θ as:

$$\nabla_\Theta\big[J\big] = \Big[\frac{\partial J}{\partial W^{(0)}}, \frac{\partial J}{\partial b^{(0)}}, \frac{\partial J}{\partial W^{(1)}}, \frac{\partial J}{\partial b^{(1)}}, \ldots, \frac{\partial J}{\partial W^{(L-1)}}, \frac{\partial J}{\partial b^{(L-1)}}\Big]^T \tag{25}$$

Where $\frac{\partial J}{\partial W^{(l)}}$ is taking the partial derivative of each element $W_{i,j}$ in the $W^{(l)}$ matrix, and preserves the shape. While there may only be 20 or so layers, the gradient vector actually has one element for every trainable parameter in the Θ object. This means that for our network, the gradient vector contains upwards of $33,000$ elements. We again choose to group these elements by their parent structure for ease of notation.

Due to the nested composition of the network output in Eq. (3), and subsequently the cost itself, we must use the chain rule of calculus to work backwards through computational graph of the neural network to compute the partial derivative of J with respect to each parameter in Θ. The process of working backwards to compute each element in the gradient vector is called *back-propagation* [2, 4, 13].

2.5.3 Back-Propagation

Recall our model of a neural network as a computational graph as in Eq. (2). Each node of the graph is a layer that represents a mathematical function which takes the activation from the previous layer $x^{(l-1)}$ and transforms them into the activations of the current layer,

as in Eq. (4). Additionally, we recall that each layer with trainable parameters is made up of two steps which produce linear activations such as in Eq. (5) and non-linear activations, as in Eq. (7). Therefore, to understand back propagation, it is helpful to examine the flow of information within each layer. Consider a simplified neural network model visualized in Fig. (5). Note that we indicate the cost as $J(\Theta)$. This is alluding to the fact that y^* is ultimately a function of the parameters in Θ.

Raw information is presented as $x^{(0)}$ at the top of the figure, and and output is returned as $x^{(L-1)}$ at the bottom. (Recall $y^* = x^{(L-1)}$) Arrows represent the flow of information in the network. Because of the nested computation nature of the graph, we use the chain rule for partial derivatives to begin at the final output of the network and successively work our way forward to the entry layer. At each step, we compute the numerical derivative with respect to the parameters in each layer. This process is called *back-propagation* [2, 4].

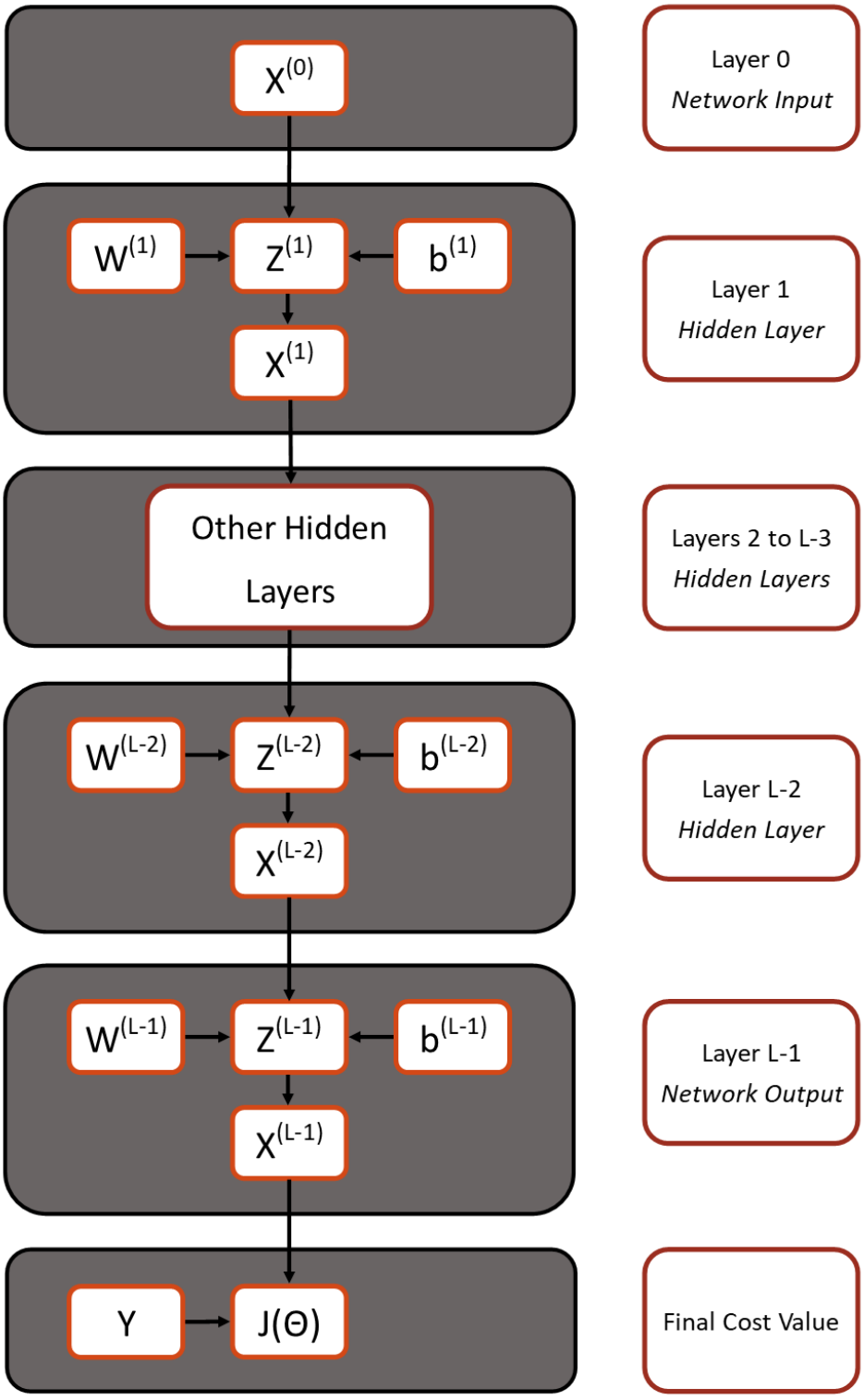


Figure 5: A visual representation of a simplified neural network as a computational graph

Suppose we wish to compute the elements of $\frac{\partial J}{\partial W^{(L-1)}}$ and $\frac{\partial J}{\partial b^{(L-1)}}$ for the gradient vector. The flow of information in Fig. (5) indicates that we cannot directly evaluate the derivative because the cost J, is a function $x^{(L-1)}$, which is a function of $z^{(L-1)}$, which is a function of elements in $W^{(L-1)}$. Thus, we have:

$$\frac{\partial J}{\partial W^{(L-1)}} = \frac{\partial}{\partial W^{(L-1)}}\left[J\left\{x^{(L-1)}\left[z^{(L-1)}\left\{W^{(L-1)}\right\}\right]\right\}\right] \tag{26}$$

The evaluation of the derivative of a composite function requires use of the chain rule of multivariate calculus. We represent this symbolically as:

$$\frac{\partial J}{\partial W^{(L-1)}} = \frac{\partial J}{\partial x^{(L-1)}}\frac{\partial x^{(L-1)}}{\partial z^{(L-1)}}\frac{\partial z^{(L-1)}}{\partial W^{(L-1)}} \tag{27}$$

Similarly, the partial derivative of the cost with respect to the bias array can be formulated as:

$$\frac{\partial J}{\partial b^{(L-1)}} = \frac{\partial J}{\partial x^{(L-1)}}\frac{\partial x^{(L-1)}}{\partial z^{(L-1)}}\frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

For the categorical cross entropy cost function, we can evaluate $\partial J/\partial x^{(L-1)}$ over a single sample that belongs to class i as:

$$\frac{\partial J}{\partial x^{(L-1)}} = \frac{\partial J}{\partial y^*} = \frac{\partial}{\partial y^*}\Big[-y\ln(y^*)\Big] = -\frac{y}{y_i^* + \delta} = -\frac{1}{y_i^* + \delta} \tag{29}$$

We often introduce $\delta \approx 10^{-8}$ to prevent possible division by 0 errors. The value $\partial x^{(L-1)}/\partial z^{(L-1)}$ can be found by differentiating Eq. (7) with respect to the input activations.

$$\frac{\partial x^{(L-1)}}{\partial z^{(L-1)}} = \sigma^{'(l)}\big[z^{(L-1)}\big] \tag{30}$$

For a dense layer as in section (2.3.1) or a 2D-convolution layer as in section (2.3.2) $\partial z^{(L-1)}/\partial W^{(L-1)}$ is given by the derivative of Eq. (5) or Eq. (6) with respect to $W^{(l)}$:

$$\frac{\partial z^{(L-1)}}{\partial W^{(L-1)}} = \frac{\partial}{W^{(L-1)}}\Big[W^{(L-1)}x^{(L-2)} + b^{(L-1)}\Big] = x^{(L-2)} \tag{31}$$

$$\frac{\partial z^{(L-1)}}{\partial W^{(L-1)}} = \frac{\partial}{W^{(L-1)}}\Big[\big(W^{(L-1)} * x^{(L-2)}\big) + b^{(L-1)}\Big] = x^{(L-2)} \tag{32}$$

Finally, differentiating with respect to the bias array: $\partial z^{(L-1)}/\partial b^{(L-1)}$, we have:

$$\frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial}{W^{(L-1)}}\Big[W^{(L-1)}x^{(L-2)} + b^{(L-1)}\Big] = 1 \tag{33}$$

Suppose we then wish to compute the elements of $\frac{\partial J}{\partial W^{(l)}}$ and $\frac{\partial J}{\partial b^{(l)}}$ for any layer (l). Using the same graph from Fig. (5), we can follow the chain of information to derive an expression for the similar to Eq. (27) and Eq. (28) for each layer (l) in the nueral network. In general, we can state that:

$$\nabla_{W^{(l)}} J = \frac{\partial J}{\partial W^{(l)}} = \frac{\partial J}{\partial x^{(l)}} \odot \partial\sigma^{(l)}[z^{(l)}] \cdot x^{(l-1)} \tag{34}$$

And

$$\nabla_{b^{(l)}} J = \frac{\partial J}{\partial b^{(l)}} = \frac{\partial J}{\partial x^{(l)}} \odot \partial\sigma^{(l)}[z^{(l)}] \tag{35}$$

Where $z^{(l)}$ are the linear activations defined in Eq. (5) and Eq. (6). Note that each $z^{(l)}$ and $x^{(l)}$ are typically stored in memory during the forward pass in Alg. (1) to prevent the need to recompute them [3, 4].

Algorithm 5 Backwards propagation system, in a standard densely connected deep neural network. Each iteration in the *for-loop* computes the gradient of the cost function J with respect to the weight and bias arrays in a given layer (l). Each element in those arrays dW and db is the discrete gradient of the cost due to that parameter. A practical application of this algorithm should include batches of samples instead of a single sample and a regularizing function at each step.

Require: Cost/Objective function J.
Require: Set of Layer functions $f^{(i)}$. Each one contains a weighting array $W^{(i)}$, a bias array $b^{(i)}$, and activation function $\sigma^{(i)}$, the activation function derivative, $\sigma'^{(i)}$. In all cases, $i \in \{0, 1, ..., L-1\}$
Require: Set of linear and non-linear activation arrays $z^{(i)}$ and $x^{(i)}$.
  Execute forward pass in algorithm (1) and compute the gradient of the cost with respect to the final layer activations
  $dx \leftarrow \nabla_{(y^*)} J(y, y^*)$
  Initialize ∇J as output object, should have same shape as Θ
  for $l \in [L-1, L-2, ..., 2, 1]$ **do**
    Compute gradient w.r.t pre-nonlinear activation portion of layer function
    $dx^{(l)} \leftarrow \nabla_{Z^{(l)}} J = dx^{(l)} \odot \partial\sigma^{(l)}[Z^{(l)}]$
    Compute gradient w.r.t weighting and bias elements
    $db \leftarrow \nabla_{b^{(l)}} J = dx^{(l)}$
    $dW \leftarrow \nabla_{W^{(l)}} J = dx^{(l)} \cdot X^{(l-1)}$
    Add db and dW steps to ∇J object
    $\nabla J = \nabla J.Add(dW, db)$
  end for
  Return gradient w.r.t to each parameter in Θ
  return ∇J

After (i) computing the gradient, we can scale it by a desired learning rate α wich controls the size of the gradient step in each component, and (ii) add the gradient vector element-wise to the existing elements in Θ object. By repeating steps (i) and (ii) in succession, we

gradually drive the cost function to produces consistently lower and lower values across a data set [2]. This is called *gradient descent* and is the basis for many optimization algorithms. Let $\bar{J}$ be the average cost over a batch of b samples. We show the general update rule on iteration (i) for gradient based learning in Eq. (36) [2, 4].

$$\Theta^{(i)} = \Theta^{(i-1)} + (-\alpha)\nabla_{\Theta}\bar{J} \tag{36}$$

Where $\Theta^{(i)}$ will give the set of new, updated parameters of the network. Note that for Eq. (36), all operations are applied element-wise.

2.5.4 The Optimizer

An optimizer is the algorithm or procedure that is used by a machine learning model to perform the optimization task to reduce the cost value after reciveing each new training sample. Regression models may often used a *mean-squared error* cost function which can often be solved analytically, while high dimensional neural networks, such as the one used in this project, outlined in section (2.6), can only up updated iteratively [4, 8, 13]. We use a variation of gradient descent much like the one in Eq. (36). A standard stochastic gradient-based or batch gradient descent learning algorithm can be fast, but are often prone errors such as vanishing or exploding gradients [2, 4, 13]. To combat this, model must often implement an optimizer that is more stable, and robust in it's ability to drive the cost function to a successively lower value.

For this project, we employ an *Adaptive-Moments* optimizer, also called *ADAM* for short. This is a powerful algorithm that uses an adaptive learning rate and built-in momentum parameter. ADAM records and updates an exponentially decaying average of past gradients, $s^{(i)}$, and an exponentially decaying average of past squared gradients, $r^{(i)}$ in Eq. (37) [2]. This produces a far more aggressive optimizer at a higher computational cost than standard gradient descent [4].

As the optimizer iterates through each step, the effect of past iterations will tend to "snowball". If the previous step was found to reduce the cost function by a large or small amount, the step size for the previous step will update accordingly. This enables to ADAM to overcome smaller discontinuities that may arise in the model's solution space [4]. For a given step (i), The ADAM update is given:

$$\begin{aligned}
s^{(i)} &= \rho_1 s^{(i-1)} + (1-\rho_1)\nabla_{\Theta}\bar{J} \\
r^{(i)} &= \rho_2 r^{(i-1)} + (1-\rho_2)\Big[\nabla_{\Theta}\bar{J} \odot \nabla_{\Theta}\bar{J}\Big] \\
s'^{(i)} &= \frac{s^{(i)}}{1-\rho_1^i} \\
r'^{(i)} &= \frac{r^{(i)}}{1-\rho_2^i} \\
\Theta^{(i)} &= \Theta^{(i-1)} + (-\alpha)\frac{s'^{(i)}}{\sqrt{r'^{(i)}} + \delta}
\end{aligned} \tag{37}$$

Note that a superscript (i) or $(i-1)$ gives an iteration index, while the superscript i means to raise a value to the power of i. ADAM has experimentally shown to be a very powerful and robust optimizer useful for a wide range of tasks [4]. Because of the two decay constants ρ_1 and ρ_2, we can compound and accumulate the values of past gradients to continue to push the cost to lower and lower values, even if the magnitude of the gradient becomes very small [2].

Algorithm 6 Adaptive-Moments (ADAM) optimizer for a neural network. This algorithm is adapted from Goodfellow [4]

Require: Step size α
Require: Small constant δ for numerical stabilization, usually about 10^{-7}.
Require: Constants ρ_1, ρ_2 used track exponential decay rates, usually 0.9 and 0.999 respectively.
Require: Subroutine/function to compute gradient of cost function See Alg. (5)
Require: Mini-batch size, m
Require: Stopping criterion S
Initialize moment variables and iteration counter $s=0$, $r=0$, $i=0$
while Stopping Criterion S is **false do**
Extract a mini-batch of m samples from larger data set X. $[x^{(0)}, x^{(1)}, ..., x^{(m-1)}]$ and corresponding target values $[y^{(0)}, y^{(1)}, ..., y^{(m-1)}]$.
Compute numerical gradient estimate of each sample in batch. This can be done with standard back-propagation in algorithm (5) and normalize by batch size m:
$\nabla \bar{J} \leftarrow \frac{1}{m}\nabla_{\Theta}\Big[\sum_{n=1}^{m} J(y^{*(n)}, y^{(n)})\Big]$
Compute first bias moment: $s \leftarrow \rho_1 s + (1-\rho_1)\nabla \bar{J}$
Compute second bias moment: $r \leftarrow \rho_2 r + (1-\rho_2)\big(\nabla \bar{J} \odot \nabla \bar{J}\big)$
First bias correction: $s' \leftarrow \frac{s}{1-\rho_1^i}$
Second bias correction: $r' \leftarrow \frac{r}{1-\rho_2^i}$
Compute And Apply update: $\Delta\Theta \leftarrow (-\alpha)\frac{s'}{\sqrt{r'}+\delta}$
$\Theta \leftarrow \Theta + \Delta\Theta$
Update Iteration number: $i \leftarrow i+1$
end while

2.6 Multimodal Architecture

In addition to choosing a strong set of features, and an appropriate optimizer for a classification task, it is also important to combine the predictors with a complementary network architecture. Since a neural network is a computational graph with each layer represented by a node, it can take on a nearly infinite number of different shapes and structures [4, 32]. The structure of the network defines the composition of it's function, F^* as in Eq.(2), and thus affects it's performance in completing any given task [2]. The layer structure of the neural network is referred to as it's *architecture*, and define it's *hypothesis - space* [4].

For this project, we have derived features that describe that audio file using two different *modalities*. A modality is a method of describing and interpreting a sample of data [18]. For example humans can be said to experience the world through five modalities - those being our cardinal senses of sight, hearing, taste, smell and touch. The sight and sound of a guitar are two very different ways of experiencing the same thing. One or the other is often enough to identify the guitar - but both together yield an even higher predictive confidence. Since modalities are *different* by nature, we cannot simply combine them into a single object and present the neural network with it. We must account for the difference by providing two distinct input layers. Modalities of seeing and hearing are captured by two different senses but ultimately processed by the same brain to produce a single prediction.

We can carry this idea over by producing a *multimodal* neural network that accepts two different forms of input for each sample. One input is an $N' \times k$ spectrogram matrix, explored in section (4.2) and the other input is a $1 \times p$ feature vector explored in section (4.3) and section (4.4). The spectrogram acts to provide an energy distribution of a signal as a function of frequency and time, while the feature vector provides qualities in a more list-like fashion. This type of supervised learning is referred to as *multi-representation learning*, and we detail a conceptual diagram of the architecture in Fig. (6) [11, 18].

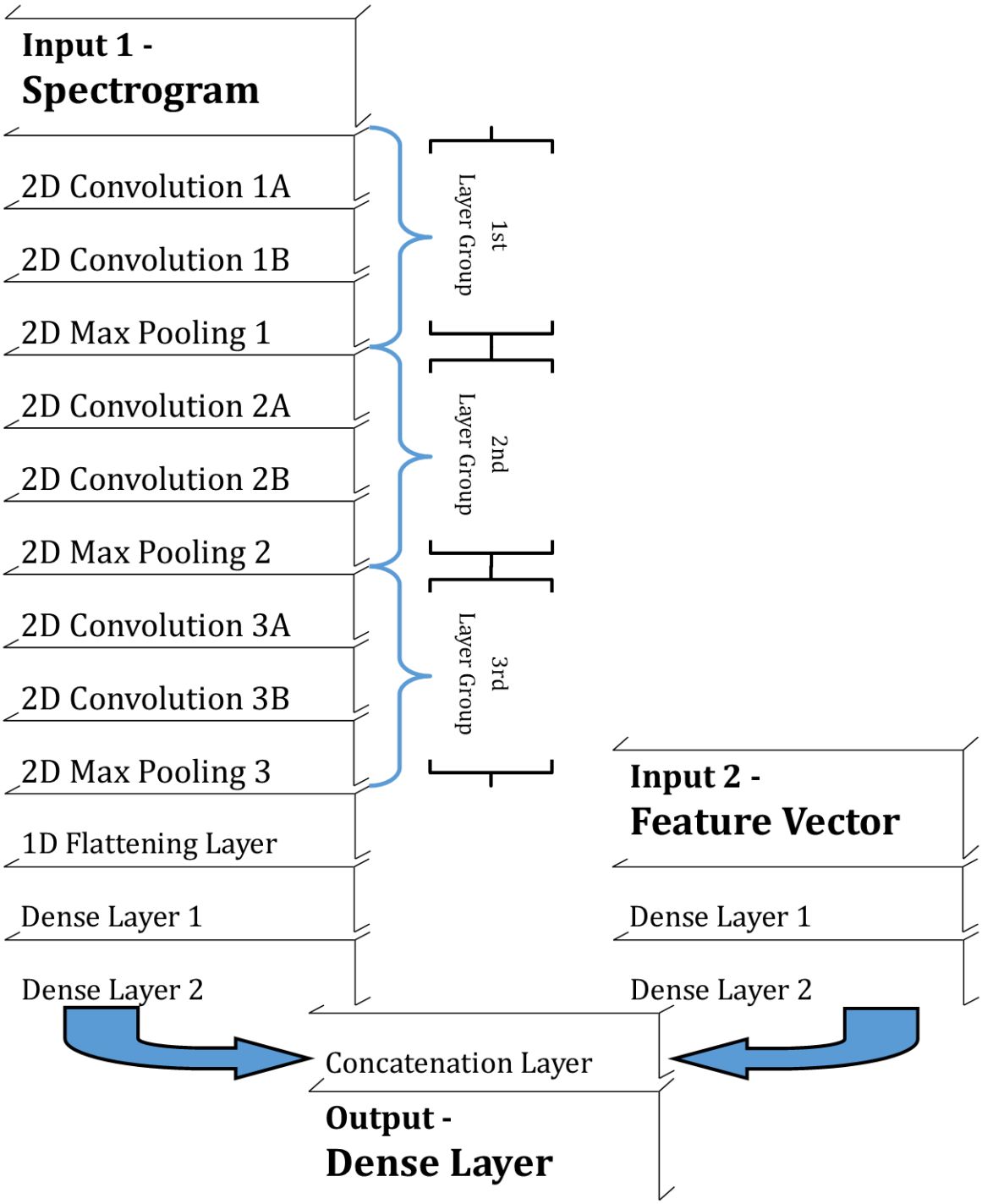


Figure 6: The implemented multimodal architecture of the audio file classification neural network. The left branch process an image-like input, and the right branch processes a vector-like input. The activations are then merged, and then a single output vector is produced

2.6.1 The Convolution Branch

The convolution branch is pictured on the left side of Fig. (6), and is an implementation of a *convolutional neural network* (CNN). It processes a spectrogram, which is a representation of a waveform, or signal as an energy distributions over *time* and *frequency* [34, 20, 9]. The input layer of this branch accepts a 4-Dimensional array. The axes, in order of indexing, represents (i) the size of the *mini-batch* of samples, (ii) the pixel height of each sample, (iii) the pixel wifth of each sample, and (iv) the number of channels in the image. For this model, we have selected to use 64 samples per batch, 558 pixels in height, 256 pixels in width, and 1 gray-scale channel. We denote the 4D shape of the design matrix for this branch as:

$$X_1 \in \mathbb{R}^{(64\times 558\times 256\times 1)} \tag{38}$$

Any other shape will be rejected, and an error is raised.

The input layer holds the input activations, X_1, which is a collection of 64 spectrograms. The output of the the input layer is passed into the first of three *Convolution Layer Groups*. These layer groups are inspired from the *Visual Geometry Group-16 Neural Network* architecture [4, 13]. Each convolution layer group is composed of three individual layers:

1. A 2-Dimensional Convolution layer, 32 filters, 3×3 kernel, 1×1 step size, ReLU activation function,
2. A 2-Dimensional Convolution layer, 32 filters, 3×3 kernel, 1×1 step size, ReLU activation function,
3. A 2-Dimensional Maximum Pooling layer, 3×3 pooling size, Identity activation function

The convolution layers iterate over the middle two axes of the data, see section (2.3.2). Physically, this allows us to search through the frequency and time axes to detect dominant features or shapes.

By grouping layers in this structure, we use convolution to reduce the number of features, and then the pooling layer to extract only the largest activations of the remaining values. This drastically reduces the width of each layer before passing the activations down, and ensures that only the pertinent characteristic shapes are preserved and passed into the next layer set for processing. This also allows for positional invariance in features. Regardless of where in frequency or time a feature is detected, the moving kernel will still allow it to be found [4, 13].

2.6.2 The Perceptron Branch

The perceptron branch is pictured on the right side of Fig. (6) and is an implementation of a *multilayer percepton neural network* (MLP). Rather than accept an image-like input, the perception simply takes a p-dimensional vector-like input of properties that are derived from the audio file. We call these properties *features* or *predictors* [2, 9, 26]. Details on the natures and selections of these p elements are found in section (4.3) and section (4.4).

The input layer of the perceptron accepts a 2-Dimensional array. The axes, in order of indexing, represent (i) the size of the *mini-batch* of samples, (ii) the number of features for each sample. We use the same model hyper-parameter of $b = 64$ samples per batch, and have developed $p = 24$ unique classification features that have been derived from time-space and frequency-space representations of the audio file data. We can denote the 2D shape if the input object into this branch as:

$$X_2 \in \mathbb{R}^{(64\times 24)} \tag{39}$$

This 2D array is referred to as a *design matrix* [8, 13]. Any other shape will be rejected by the model, and an error is raised.

In perceptron models, the scaling of the design matrix is vital to classification performance. A design matrix is scaled by taking all samples in a column (a particular feature from each class), subtracting the average, and then scaling it such that it has unit variance [2, 8]. This ensures that no one feature dominates the performance of the model and that layer activations do not get saturated to extremes as data progresses through the network. This operation was implemented through the *scikit-learn* Python package [29].

2.6.3 The Final Output Branch

The last layer in each of the two branches is a ReLU-activated Dense layer containing 64 neurons, represented by a 64 - dimensional vector. We combine these model layers by using a *concatenation layer* to fuse the arrays together, such that each sample is given by a 128-dimensional vector. This vector is then transformed into the final dense layer, which uses the softmax activation, and encodes the joint predictions based on contributions of both model branches. In section (6) we explore how the predictive power of the network changes when examining the performance of each individual branch compared to the hybrid model.

3 Properties of Musical Instruments

As discussed in section (2), performance quality of a neural network is greatly dependent on the properties of the chosen features [32, 12]. To ensure that the classifier model performs adequately and consistently, we must choose these features to have high variance between classes, and low variance within each class. This enables a neural network to development clear decision boundaries between each unique class [8, 26]. To develop this set of features, we explore the physical and mechanical properties of musical instruments and the sounds that they create.

Since computer algorithms such are neural networks are built to handle exclusively quantitative information, we must ensure that each property can be represented by a numerical quantity [2, 10]. Where a human would use qualitative adjectives such as *bright* or *harsh* or *percussive* to describe sound, computers require a numerical descriptor instead. The goal of this section is to develop the attributes of sound that enable us to describe a sound wave in a quantitative and compact way. Thus we use values that can be expressed as numbers such as *amplitude envelopes*, *frequency spectra*, or *formant structure*. For this section, we represent waveforms as *spectrograms* to capture the time-space and frequency-space development of energy simultaneously. We motivate and describe the production of a spectrogram in section (4.2).

We have assembled training data samples than can be grouped into 37 classes of unique sources listed below:

Class Index	Class Name	Counts	Class Index	Class Name	Counts
0	Alto Flute	72	20	Marimba	364
1	Alto Saxophone	128	21	Oboe	666
2	Banjo	74	22	Sawtooth Wave	200
3	Bass	1060	23	Tenor Saxophone	732
4	Bass Clarinet	1036	24	Sine Wave	200
5	Bass Flute	76	25	Soprano Saxophone	128
6	Bassoon	800	26	Square Wave	200
7	Bass Trombone	54	27	Tenor Trombone	66
8	$B\flat$ Clarinet	938	28	Triangle Wave	200
9	Bells	164	29	Trombone	831
10	Violoncello	1079	30	$B\flat$ Trumpet	627
11	Contrabassoon	710	31	Tuba	1046
12	Crotale	50	32	Vibraphone	334
13	$E\flat$ Clarinet	78	33	Viola	1173
14	English Horn	1382	34	Violin	787
15	Flute	1032	35	Whitenoise	200
16	Guitar	106	36	Xylophone	176
17	Hihat	10	37		
18	French Horn	740	38		
19	Mandolin	364	39		

Figure 7: Instruments categories used in the classification task. Note that the network uses only integers given by "Class index" to identify sources. The string "Class Name" is kept for human readability

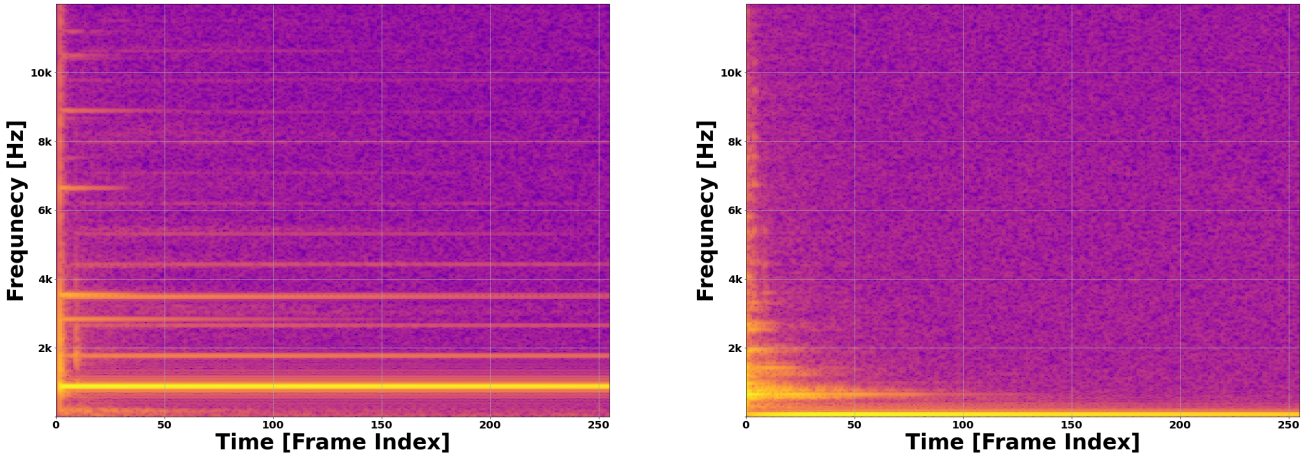
Rather than explore each class individually, we can examine the properties of groupings of musical instruments by using the *Hornbostel-Sachs* organization system developed by Erich M. von Hornbostel and Curt Sachs, and published in 1961 [6]. In this system, musical instruments or sound sources can be divided up into four broad categories based on the nature of the sound-producing material. These categories are (i) idiophones, (ii) membranophones, (iii) chordophones, and (iv) aerophones. In this section, we show how the physics of the source instrument influences some of our feature choices.

3.1 Idiophones

An idiophone is an instrument that produces sound through the vibration of the full body of the object. This generally includes most percussive instruments excluding drums [6]. From the set of classes that we use, bells, crotales, Hi-hats, marimbas, vibraphones, and xylophones are all examples of idiophones. In each case, sound comes from the vibration of the object itself or some subset of the body. In the case of mallet percussion, the wood or metal keys

vibrate when struck with a felt or yarn mallet. Sometimes the vibration is amplified by nearby air columns, but it is the vibration of the key itself where the sound originates from [20].

Idiophones in particular can be characterized by their *transient response* [34]. The transient response is the time-evolution response of the waveform in time, which we normally break into *attack*, *decay*, *sustain*, and *release*. For most idiophones, the system is struck once and left to a free or dampened vibration; which we can mathematically model as a rigid body reacting to an impulse [7, 20].



(a) Vibraphone playing $A5$

(b) Marimba playing $C2$

Figure 8: Spectrograms from idiophone waveforms

From the spectrogram representation of idophone waveforms , we can see the majority of energy in the time domain is right at the start of the waveform. This is shown by the vertical bright yellow line at on the left side of each image in Fig. (8). When an idiophone is struck, all of the energy for the vibration is delivered up front, and the amplitude can only decay. This gives the resulting transient a very short attack time, followed by a short decay, sustain, and often long release as the last of the energy dies out [34]. While other instruments such as plucked strings like guitars show similar response, this is a very important property of many idiophones. Notice that the energy of the higher overtones die out very quickly in each spectrogram, and only the lowest pitch - the fundamental is preserved for much of the duration.

This property allows for us to distinguish idiophones from other instruments by an approximation of their amplitude envelope. For example, if we divide the time series waveform into q distinct sections, and compute the RMS energy in each section, we should see a gradual decay of energy with time. More specifically, the energy in some section q_i should be greater than the energy in any subsequent section $q_{i+n}, n \in \mathbb{Z}$. Having this property does not guarantee the source is an idiophone but can we can state within reason that most idiophones

obey this behavior. In addition to the amplitude envelope, we can consider the distribution of time-space energy in the form of a single scalar quantity as the center of mass. If we treat the waveform as a 1D mass distribution, and compute the center of mass, we should see consistently low values because the largest values of amplitudes happen early on in the time-domain.

3.2 Membranophones

A membranophone is an instrument that produces sound through the vibration of a tight membrane [6]. This includes most drum-like instruments While none of the instruments that we classify are membranophones, they do exhibit some noteworthy properties that influence our feature for other instruments. Common examples include the tympani, the bass drum, or the kettle drum. Each of these instruments is composed of a large cylinder or bowl-like structure that serves as an amplifying cavity [20, 34]. Given that most membranophone instruments are played by striking the membrane with the hand or mallet, the amplitude envelope properties

We model the tight elastic membrane very similarly as that of a tight elastic string, except in two dimensions [34]. Unlike the chordophone section (3.3), the partials for the membrane are not related by integer multiples. This lack of harmonic structure leaves most membranophones to have a sort of "muddled" frequency spectrum where we see a more continuous distribution of constituent frequencies arise as opposed to clean integer related overtones [34, 20]. This partly why many membranophones have a much more "percussive" sound as opposed to performing any distinct note.

3.3 Chordophones

A chordophone is a musical instrument that produces sound through the vibration of strings that are stretched between two fixed points [6]. From our data set, banjos, basses, cellos, guitars, mandolins, violins, and violas are all examples of chordophones. We commonly model a chordophone as having a string of length L, with a constant linear mass density, μ, subject to a tension force F_T. The string is flexible, and free to vibrate in a spatial single dimension, x. We describe the state of string over space and time with a function $U(x,t)$. The behavior of the string can be modeled by D'Alembert's 1-dimensional wave equation [5, 7, 30]:

$$\frac{\partial^2 U}{\partial t^2} = v^2 \frac{\partial^2 U}{\partial x^2} \tag{40}$$

Where v is the wave propagation velocity, given by $\sqrt{\frac{F_T}{\mu}}$. D'Alembert's equation indicates that the acceleration of any point on a string is directly proportional to the rate of change of the slope of the string at that point [30].

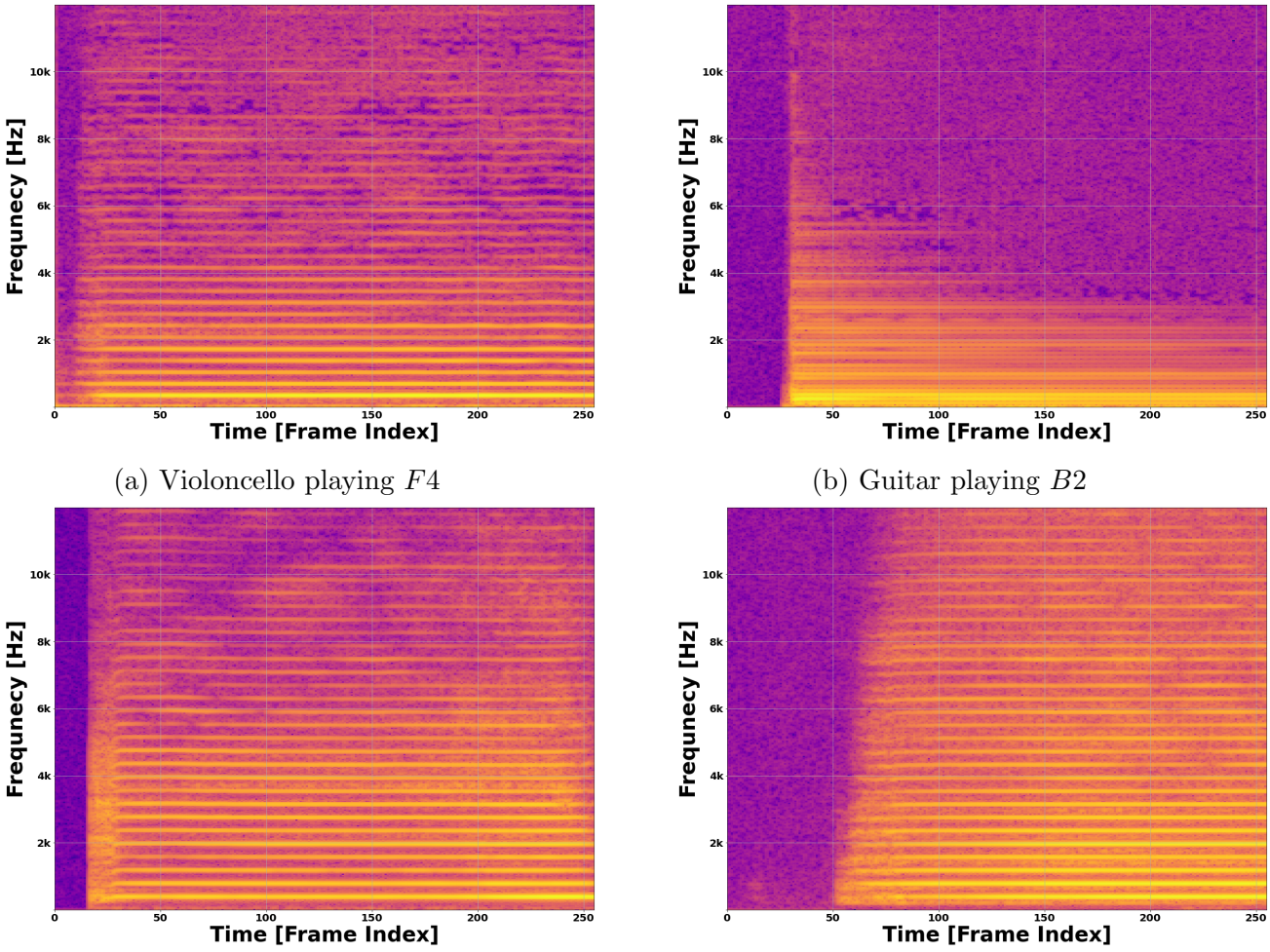
Since the PDE is second order in time, we must also specify two initial conditions being the shape and the derivative of the shape at time $t = 0$. [20, 34, 5]:

$$U(x,0) = f(x) \qquad \text{and} \qquad \frac{\partial U}{\partial t}(x,0) = g(x) \tag{41}$$

Additionally, the string is fixed in place at both ends, with a nut at $x = 0$ and a bridge at $x = L$, which can be modeled by the boundary conditions

$$U(0,t) = 0 \qquad \text{and} \qquad U(L,t) = 0 \tag{42}$$

Depending on the exact nature of the source instrument, $f(x)$ and $g(x)$ can take on different forms. For example, for plucked stringed instruments, $f(x)$ may appear to be as triangle-like function, with a peak at some position $0 < x_0 < L$ where then string was struck. It is these initial conditions that allow for the drastically different timbres that arise in the wave forms [7]. Compare the spectrogram representation of a few stringed instruments in Fig. (9).



(a) Violoncello playing $F4$

(b) Guitar playing $B2$

(c) Violin playing $G4$

(d) Viola playing $G4$

Figure 9: Spectrograms from chordophone waveforms

The wave equation is a *linear* partial differential equation which allows for a linear combination of solutions to form a sound wave. In the case of the taught finite string that is fixed at both ends, such a guitar, viola or cello, the a solution tends to take the form of a sine or cosine function [7, 30, 34]. Notice the terms arise by the many a parallel yellow and orange lines in each spectrogram in Fig. (**??**). The verical position of any line is given by the linear frequency of the sinusoid, and the brightness is proportional to the coefficient of the combination. The boundary conditions, and most forms of standard initial conditions allow for the each sinusoidal to fit exactly an integer or half-integer number wavelengths into the region $0 \leq x \leq L$ [5, 7]. Players of chordophones change pitch by shortening the length of the vibrating string or applying additional tension to the string. This changes the length of portion of the string that vibrates, and causes the production of a new pitch.

Each chordophone instrument does produce a slightly different frequency spectrum, hence each provides a slightly different timbre to the listener, so differentiating between them is possible. Because of this, we find that the spectrogram representation of the waveform is so critical to classification success because the convolution layers allow for the detection of small variations in the frequency-time domain where humans would otherwise be unable to [13, 32]. However, the size of the instrument limits the length of the string that vibrates which leads to limitations in the range of the overtones the instruments will produce. For example, a cello playing a low $C1$ will have overtones extending up into the 12 kHz range, but the energy will be very low there, thus have a very low frequency center of mass. Compare this with a violin playing a $C6$, will have overtones into the 20 kHz range, and will produce a very high frequency center of mass. While the ranges of the musical instruments overlap a great deal, the energy distribution in frequency space can be treated as a 1D mass distribution, and the frequency center of masses of instruments with similar timbres can still differ greatly [26].

Notice as well in the spectrograms of Fig. (9), how the transient response (the time domain) of chordophones differs wildly from idiophones and membranophones. In the case of bowed instruments, the attack of the stringed instrument is much longer, with a much more gradual sustain decay, and release. Where an idiophone or plucked string has all of the time-domain energy up front, the bowed string waveform continues to build energy as the instrument is bowed, and only loses energy once the bowing stops, and the string is left to vibrate by itself [7]. This causes a bowed stringed instruments to present a very different amplitude envelope compared to the other categories of instruments, and a much more centrally located temporal center of mass.

3.4 Aerophones

An aerophone is a musical instrument that produces sounds through the vibration of an air molecules in a column-like structure. This is the largest category of instruments that we have, and is commonly broken down into smaller subcategories such as *woodwind* and *brass* instruments. From our data set - alto flutes, alto saxophones, bass clarinets, bass flutes, bassoons, $B\flat$ clarinets, contrabassoons, $E\flat$ clarinets, English horns, flutes, oboes, tenor

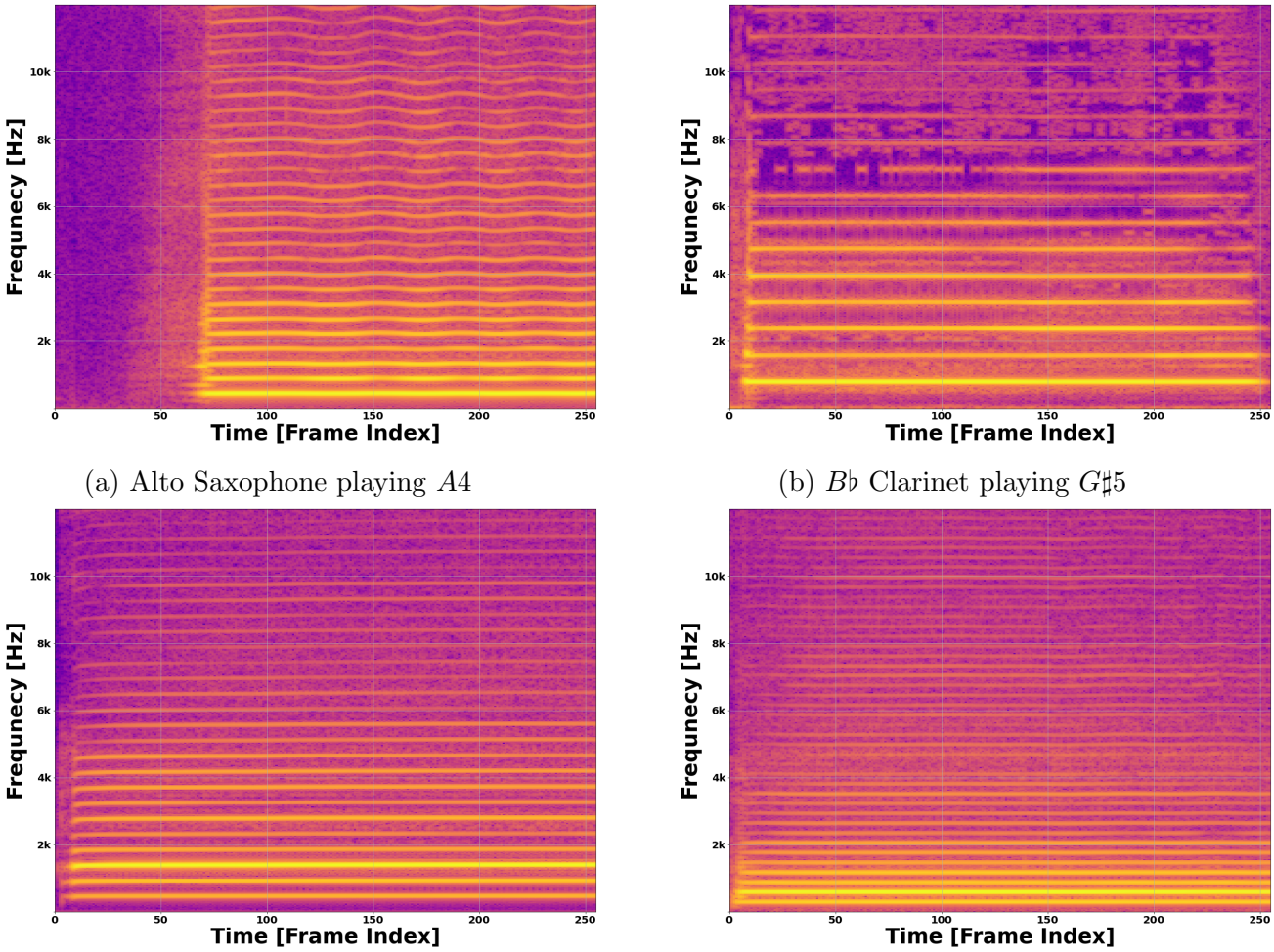
saxophones, and soprano saxophones are all examples of woodwind aerophones. Additionally, bass trombones, french horns, tenor trombones, trombones, $B\flat$ trumpets, and tubas are all examples of brass aerophones [6, 20, 34].

Aerophones all produce sound by exciting molecules of air within a column that vibrates in different patterns called *modes* [34]. the wavelength and energy of each mode is determined by the structure of the instruments, and the actions of the player. Most aerophones such a flutes, clarinets, and saxophones are more much more reminiscent of air columns that are closed at one end and open at the other. This is marked by the presence of a mouth piece at one end where a played blows air from their lungs to add energy, and a bell at the other end to serve as an amplification structure [20]. For a column of length L, that is open at one end, the fundamental frequency of the vibration is given by [20]:

$$f_0 = \frac{v}{4L} = \frac{v}{\lambda} \tag{43}$$

where v is the velocity of sound in the medium, and λ is the wavelength of the sound wave. Eq. ([**?**]) gives the lowest pitch that such an instrument with a length L can play. Lower woodwinds such as bass clarinets, have a much longer vibrating air column that that of the much shorter piccolo, thus can produce considerably lower notes.

To change the pitch, the performer of an aerophone must change the length of the vibrating air column by pressing keys with their fingers. This causes a different number of wavelengths to fit within the column, and thus change the note being played. Shortening the column (reducing L) also causes a reduction in the wavelength, and an increase in frequnecy and vice-versa. In Fig. (10) we show the spectrogram waveforms for a few different woodwind instruments.



(a) Alto Saxophone playing $A4$

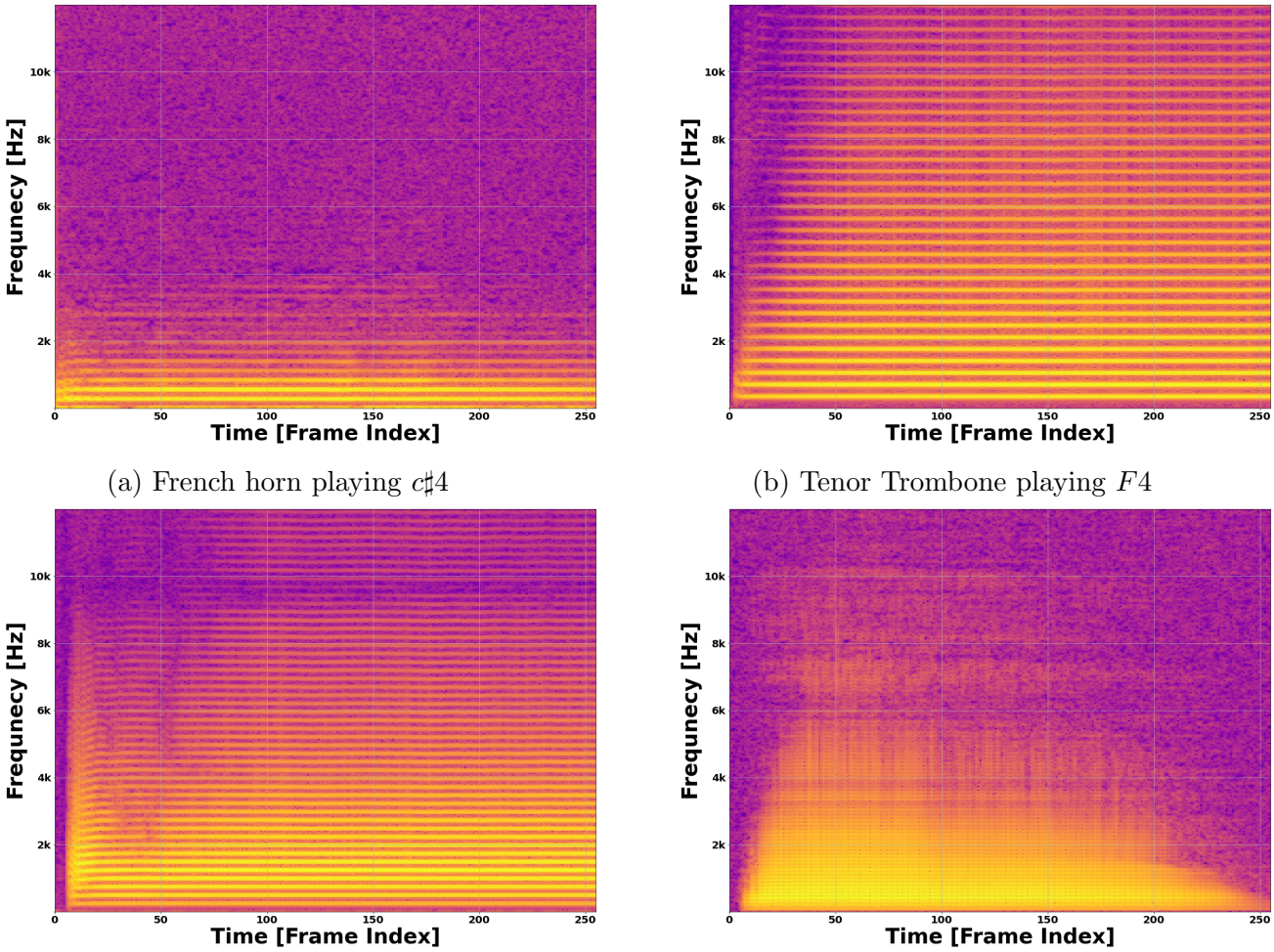
(b) $B\flat$ Clarinet playing $G\sharp 5$

(c) Oboe playing $B\flat 4$

(d) Bassoon playing $D4$

Figure 10: Spectrograms from woodwind aerophone waveforms

In addition to the woodwind instruments described above , we also consider the properties of brass aerophones. As the name implies, brass instruments are usually made of interconnected hollow tubes of brass as opposed to linear columns of mostly wood. Brass aerophones are also modeled as open-closed columns that vibrate and are also shortened and lengthened by pressing keys. Brass instruments are generally much longer than woodwinds, and produce sounds at a generally greater volume [20]. Despite these differences, brass instruments do generally display a great deal of similarities with wood winds, as shown in Fig. (11)



(a) French horn playing $c\sharp 4$

(b) Tenor Trombone playing $F4$

(c) $B\flat$ Trumpet playing $B3$

(d) Tuba playing $A2$

Figure 11: Spectrograms from brass aerophone waveforms

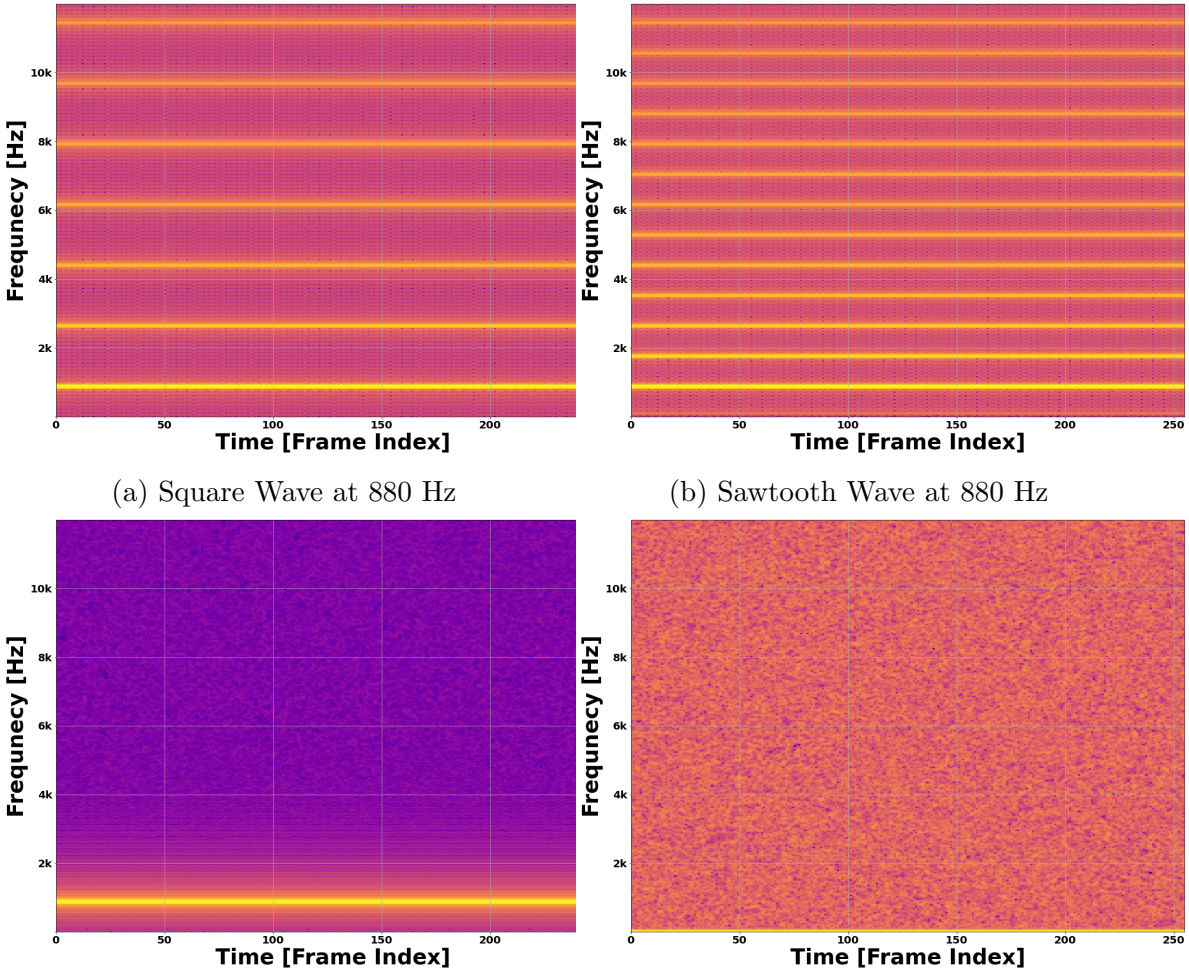
Just like with idiophones and chordophones, the amplitude envelop of a brass instrument is very characteristic. Aerophones require a user to vibrate air with their lungs which gradually adds energy to the waveform. This transfer of energy can take comparably long time, giving aerophones particularity long attack times when compared to idiophones or membranophones. The note is sustained as long as a player is vibrating the air, allowing for any length of decay and sustain, but the energy does out very suddenly when the player stops, giving a characteristically short release time [34].

3.5 Other Generated Sounds

We group sounds produced from non-physical musical instruments in this final category because they are not considered in the original Hornbostel-Sachs system. From our data set, this category includes the four simplest waveforms begin, sine waves, sawtooth waves,

square waves, and triangle waves, as well as white colored noise [34]. These sounds have been produced synthetically through a MATLAB program specifically for this project. Because of their synthetic nature, they do not display many of the physical or mechanical properties of the previous sections.

Each of the five synthetically generated waveform types were produce to have a constant envelope with all time. This means that each waveform has no characteristic amplitude shape and the spectrogram shows no time-evolution of energy as expected. Each of the the signals can be composed of a linear combinations of oscillating sinusoidal at various different frequencies and energies. For example, a sine wave consists only of a single oscillating term, while a square wave consists of repeated odd harmonics. White noise consists of a roughly uniform distribution of energy, thus shows no energy evolution in time or frequency space. These behaviors can be observed in Fig. (12).



(a) Square Wave at 880 Hz

(b) Sawtooth Wave at 880 Hz

(c) Sine Wave at 880 Hz

(d) White Noise

Figure 12: Spectrograms from synethetically generated waveforms

These waveforms were included to our data set to serve as a control to understand how the chosen features respond to simple waveforms. Additionally, the elementary composition of each waveform allows for it to potentially resemble a wide variety of instrument frequency spectra [34].

4 Feature Selections

Classification tasks require a set of inputs called *predictors* or *features* [8, 13, 26]. A feature is a quantitative, low-dimensional representation of a sample that conveys its important characteristics. For example, in classifying types of animals, a feature could be the mass, volume or number of teeth on the animal. Typically, we produce a set of p features for each sample and assemble them into a p-dimensional vector called a *feature vector* which acts as a list of descriptive qualities of the sample [2, 8]. This list is presented to the neural network and used to Audio machine learning researchers M. Kashif Saeed Khan and Wasfi G. Al-Khatib describe the vitality of feature selection [9]:

> The data reduction stage which is also called feature extraction, consists of discovering a few important facts about each class. The choice of features is critical as it greatly affects the accuracy of audio classification.

An adequate selection of features is vital to the process of training and evaluation of the classifier model [16, 26, 12]. Consider the task of identifying fish from sharks using only the presence of gills. Since both sharks and fish have gills, the features are correct representations, but not useful for discerning them. A more appropriate choice would be compare the number of gills that fish or sharks have, which would provide a correct representation and allows for differentiation.

In the biological process of categorizing sound sources, the time-domain waveform is usually enough to match the sound to a source [20]. However, a computer representation of a waveform is an array-like structure of values sampled discretely in time and volume [32, 12]. Presenting the raw time-series waveform to a model directly has experimentally shown to be unstable in for classification tasks so we must develop a set of features that can describe important properties of the waveform in a far more efficient manner [4, 8, 26]. To ensure the construction of a suitable model, we derive features based from three representations of the audio file: (i) a spectrogram matrix, (ii) the time-space, and (iii) the frequency-space. It is important to note that although this algorithm will map audio files to source instruments, it will never actually be presented with a waveform directly, and instead will rely solely on the features to make predictions.

4.1 Feature Space

From each audio file sample, we calculate the values for the same predictors. The predictors are arranged into two separated structures, each called a *design matrix*, described in section (4.1.2) [4, 2]. For a model with p predictors, feature space is a p-dimensional vector space where each basis vector represents a predictor [4, 8]. Each sample can be described by a p-dimensional vector, given by each row of the design matrix X_2. For samples $x^{(a)}$ and $x^{(b)}$ that belong to the same class, we expect the vector to be similar, $x^{(a)} \approx x^{(b)}$ in some way. This indicates that the two samples lie "close together" in space. Similarly, for samples $x^{(c)}$ and $x^{(d)}$ that belong to different classes, we expect the vectors to be very different, indicating

that the two sample lie "far apart" in space. Note that this is an idealization, and not always the defining case for the decision function behavior.

We can expand this idea by considering by how features should behave within each class, and between each class. Tuomas Virtanen, machine learning and audio engineer writes in his book, "Computational Analysis of Sound Scene and Events" about specifically how features should be selected [32]:

> For recognition algorithms, the necessary property of the acoustic features is low variability among features extracted from examples assigned to the same class, and at the same time high variability allowing distinction between features extracted from examples assigned to different classes.

This is to say, features should *ideally* form non-overlapping groups or bubbles that allows classifiers to effectively differentiate categories. Given the complexity of real-world problems, this is an often unobtainable [4, 8]. Under the right circumstances, classifiers can still produce parameters that allow for reasonable performance [2, 13]. In practice, a neural network does not rely on comparing the norms of different samples in feature space, but rather uses something called a *decision boundary*.

4.1.1 Decision Boundaries

A trained classifier makes a prediction in the form of a statistical likelihood of a sample belonging to a particular class [4, 8]. A k- categories classifier contains k output neurons, each one representing a different class. When a set of features is presented to the classifier, the value contained within each neuron is the "confidence" on the interval $[0, 1]$ that a sample belongs to that class. For example, if neuron q has an activation value of 0.25, we say that there is a 25% chance that the sample belongs to class q. Most commonly, we take the neuron with the highest activation value to be the chosen class, even if the next highest activation is a close second [13]. Thus the formal prediction function is the index of the maximum value in the output vector.

This concept stems from the idea that when a trained neural network classifier makes a prediction, it is basing class probabilities on a set of generated *decision boundaries* [2, 8]. To explore decision boundaries, consider a simplified toy data set with a 2D feature space for a classifier that must only discern between *Class A* and *Class B*. We can visualize each sample in this 2D space, and color-code each sample according according to it's class label as in Fig. (13).

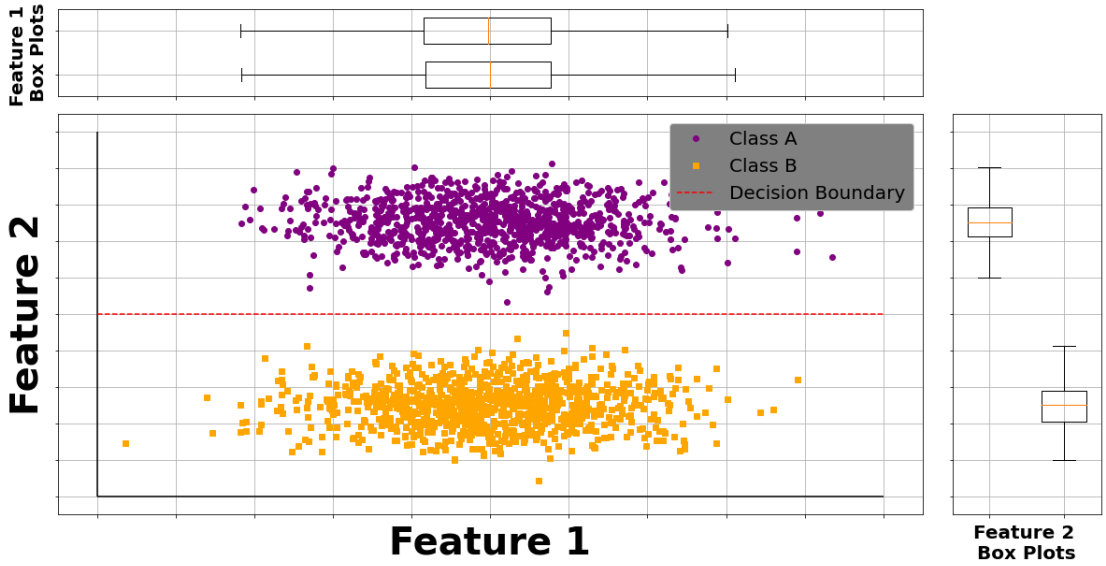


Figure 13: Feature Space with Linear separability between two classes

In the case of these chosen, features, we can see a clear, well-defined separation between samples in each class. In practice, this is rarely the case, but shows graphically an example where a linear decision boundary arises [8]. Notice how we can also use "box-and-whisker plots" to neatly summarize how each feature (each component) behaves by dividing the samples for each class based into quartiles. Notice how on the right of Fig.(13) the box plots for feature 2 show no overlap, but the box plots for feature 1, at the top show very similar behavior between classes. Despite this, a classifier can produce a clean boundary to separate the two classes. Note that this may not be the exact boundary used by a model, but it represents an ideal case.

We can also explore an example where both features have mutual overlap between classes, but we can still form a reasonable decision boundary in the space. Consider Fig. (14) and Fig. (15). In both cases, we see the box plot for each feature overlap considerably, and we also see sample that appear to "cross" the decision boundary. This crossing the boundary is why it is so useful that we choose an activation function that encode the output of the neural network as a probability of being in either class as opposed to a hard prediction [13, 2].

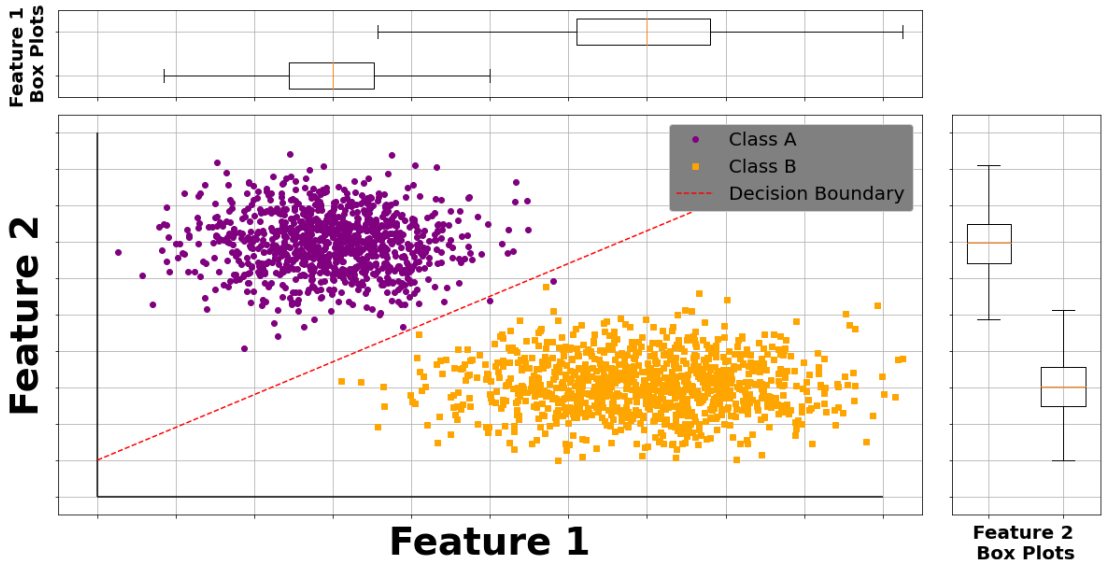


Figure 14: Feature Space with near linear separability between two classes

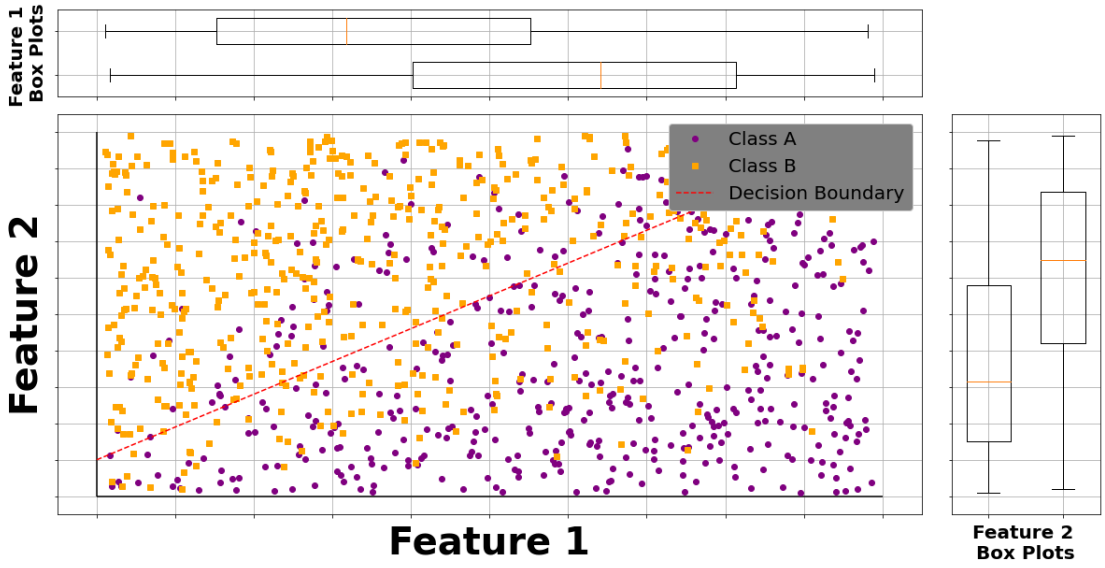


Figure 15: Feature Space with near linear non-separability between two classes

Samples that appear very near to decision boundary will likely have almost equal probability of occurring in either class, and samples that appear very far from the decision boundary will likely have very much larger probability of occurring in one class or the other [1, 8].

Note that these are over-idealized cases, constrained to two classes in two dimensions. The classifier implemented in this project uses 24 classes in a single branch and outlines a probabilty estimate over 37 classes. Given the related nature of many of the categories, (i.e. belonging to similar Hornbostel-Sachs groups), such well-defined decision boundaries are unlikely to be produced. Note also that the examples provided showed only *linear* boundaries, where the inclusion of non-linear activation functions will create far more complex systems in a far higher-dimensional space [4].

From these examples, we conclude the mathematical nature of features allow each sample to be characterized by a vector in p-dimensional feature-space [4, 8]. We expect feature of the same class to exhibit similar characteristics, and features of differing classes to exhibit different characteristics. This enables a trained classifier to develop multiple decision boundaries in the higher-dimensional space [8, 1].

4.1.2 The Design Matrix

Each audio sample is represented by an $N' \times k$ spectrogram matrix explored in section (4.2), a $1 \times p$ feature vector, and a one-hot-encoded target label, y. In order to efficiently present the information to the model for training or predicting, we extract a batch of b samples and produce the two arrays form each. Since the spectrogram and feature vector represent different *modalities*, we construct two separate *design matrices*, X_1 and X_2 [2, 11, 18]. Each has shape given by:

$$X_1 \in \mathbb{R}^{(b \times N' \times k)} \tag{44}$$

$$X_2 \in \mathbb{R}^{(b \times p)} \tag{45}$$

Where $N' \times k$ gives the shape of the spectrogram matrix S from a single sample, described in section (4.2). Similarly, p is the number of predictors extracted from each sample, outlined in section (4.3) and section(4.4). The first axis, with size b indicates that there are b audio file samples stored in that design matrix. Organizing the samples in this fashion allows for easy machine and human interpretation of each axis and index.

A design matrix is typically accompanied by a *target matrix* Y, which indicates the class that the particular sample belongs to. For a classifier with κ classes, each sample labeled by a one-hot-encoded vector for shape $1 \times \kappa$ [31, 13]. Thus, for the design matrices X_1 and X_2, we have a corresponding target matrix given by:

$$Y \in \mathbb{R}^{(b \times \kappa)} \tag{46}$$

4.1.3 Audio Preprocessing

Preprocessing a data set is a necessary step to execute prior to feature extraction [3, 8, 26]. In the case of audio files, preprocessing usually consists of ensuring that the data set contains the following steps and requirements:

1. A suitably sized number of files of reasonable audio quality with normalized amplitudes
2. Audio encoded in a standard, and consistent format
3. A consistent sample rate and bit depth between audio files
4. A consistent number of channels

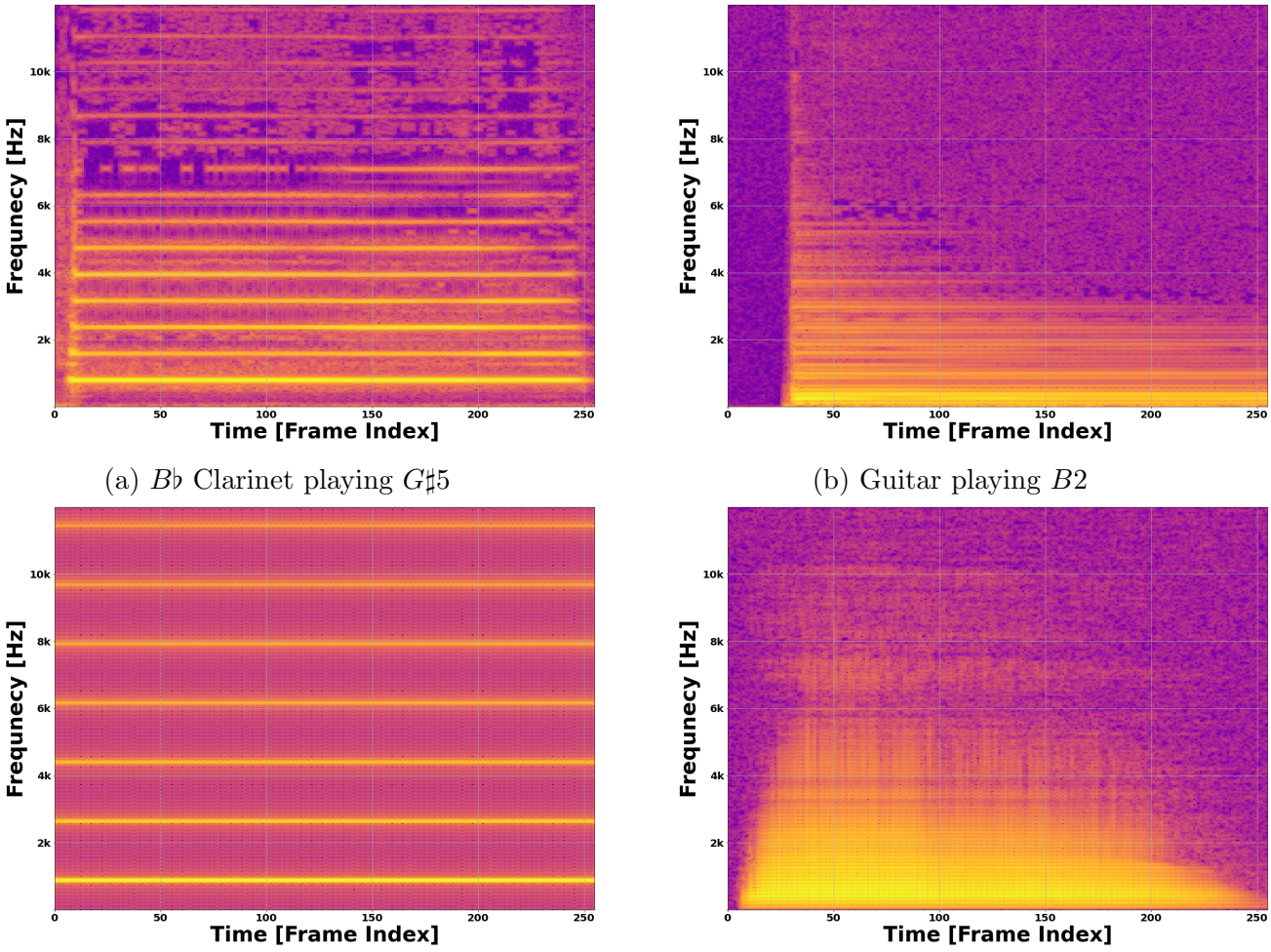
Note that different projects may require a different set of requirement from preprocessing [32]. For this project, we have chosen to use the following parameters:

1. Roughly $18,000$ audio files Professionally or semi-professionally recorded in a studio [23, 33] All ampltiudes have been normalized to ± 1 unit.
2. All audio has be converted into *.WAV* files from other formats, such as *.AIF* or *.MP3* using a MATLAB program
3. All audio is sampled at $44,100$ Hz and 24-bit depth.
4. All audio has been down-mixed into mono-channel waveforms.

Given these standardization methods, we can exact the predictors for each sample and explore how they behave mathematically, and what they represent physically.

4.2 Spectrogram Features

The application of neural networks to image-processing and recognition has been well studied and developed in the last few decades [2, 4, 13, 16]. As a result, model architectures for image processing related tasks are well-explored and have shown experimentally successful behavior. Following this success, it is reasonable to provide an image-like representation of a sound wave as a feature. We do this in the form of a *spectrogram* matrix. A spectrogram is a representation of the energy distribution of a sound wave as a function of both frequency and time. In a conventional spectrogram, the passing of time is shown along the x-axis, and the frequency spectrum is shown on the y-axis. Thus each point in the 2-Dimensional space is an energy at a given time and frequency. Examples spectrograms from the wave form data set are shown in Fig. (16).



(a) $B\flat$ Clarinet playing $G\sharp 5$

(b) Guitar playing $B2$

(c) Square Wave at 880 Hz

(d) Tuba playing $A2$

Figure 16: Spectrogram representations of various waveforms. Note that all spectrograms are color-coded according to the log-power spectrum.

4.2.1 Frame Blocking

A spectrogram is produced by the method of *frame-blocking*, which is very prevalent in audio signal processing [12, 35]. Frame-blocking takes a raw waveform or signal, s and decomposes it into a set of analysis frames, a_i, with each being a fixed N samples in length, and has a fixed overlap with the previous frame. Each of the k frames then allows for a section of the signal to be analyzed in a *quazi-stationary state* [9, 26]. Below we visualize how analysis frames are related to the full time-series waveform.

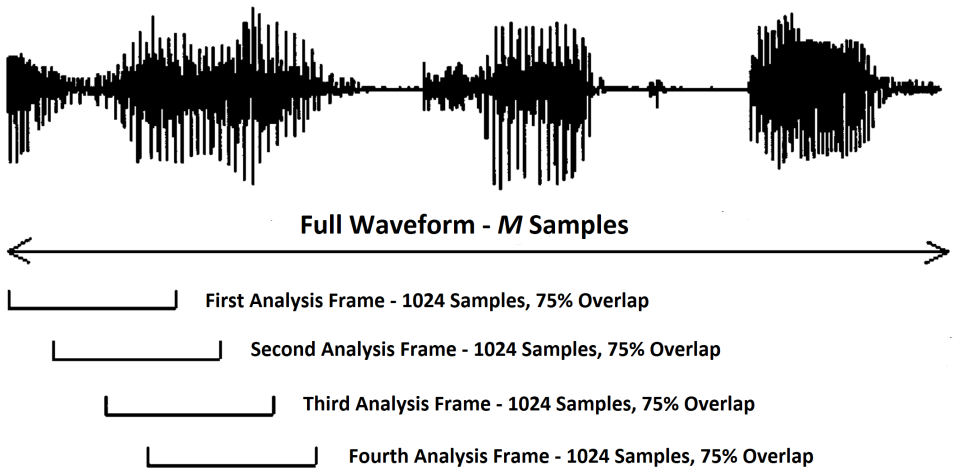


Figure 17: A visualization of how frame-blocking is used to create each analysis frames. This image has been adapted and modified from Liu, et. al. "Audio Feature Extraction and Analysis", Fig. (1). See ref. [12].

For this project, we have chosen to use frames of size $N = 1024$ with a 75% or 768 sample overlap. Since each audio file contains a different number of samples, we choose the number of frames, k to be less than or equal to 256. If $k > 256$, the waveform in truncated, if $k \leq 256$ the frames left as is, and missing frames are accounted for later (to improve computation time). The audio has been sampled at $f_s = 44100$ samples/second, so each frame represents a slice of time that is about 0.0232 seconds long.

We concatenate each analysis frame, $a_i, i \in [0, k-1]$ into a single $k \times N$ matrix, called A. Each row is a frame, each column is a sample in each frame

$$A = \{a_0, a_1, a_2, ..., a_{k-1}\} = \begin{bmatrix} a_0[0] & a_0[1] & a_0[2] & \dots & a_0[N-1] \\ a_1[0] & a_1[1] & a_1[2] & \dots & a_1[N-1] \\ a_2[0] & a_2[1] & a_2[2] & \dots & a_2[N-1] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{k-1}[0] & a_{k-1}[1] & a_{k-1}[2] & \dots & a_{k-1}[N-1] \end{bmatrix} \tag{47}$$

We use bracket notation, $a_i[j]$ to indicate that each analysis frame a_i is array-like. The following indexing conventions for matrix A all represent the same entry:

$$A_{i,j} = A_i[j] = A[i,j] \tag{48}$$

4.2.2 Windowing

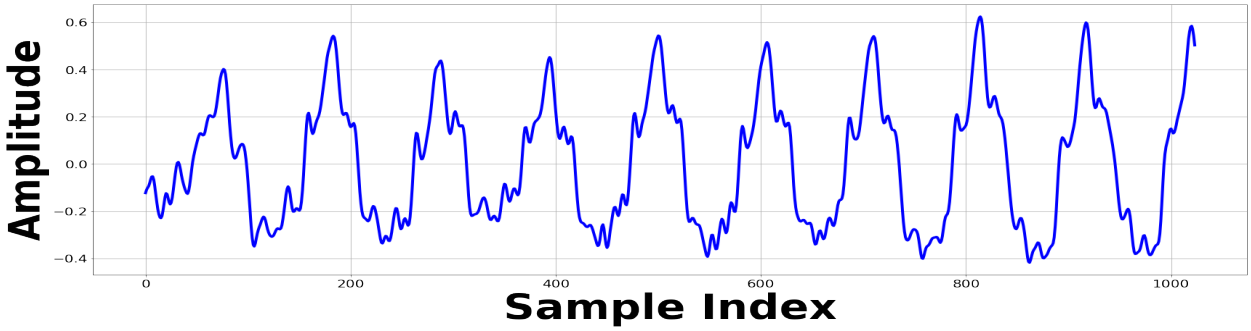
After frame-blocking, we apply a *windowing function* to each frame. A standard *Hanning Window* of N samples is generated as a $1 \times N$ row-array, H. The n-th index in a Hanning window with N samples is defined:

$$H[n] = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{49}$$

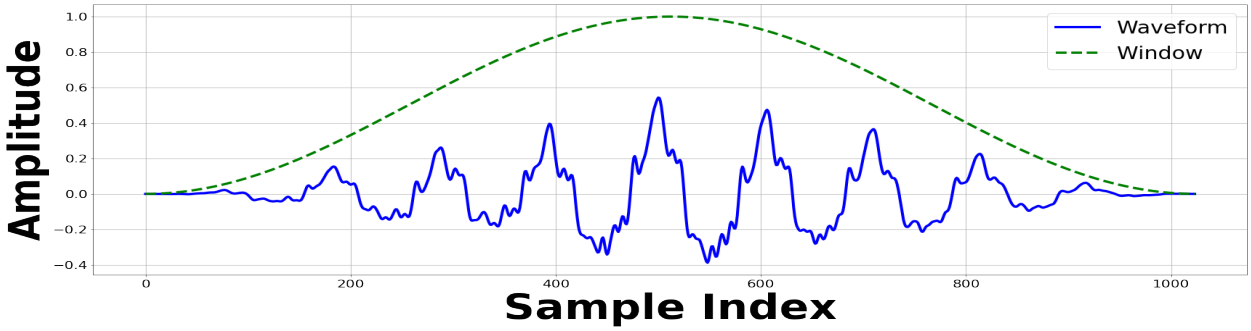
This window function is applied to each analysis frame by computing the element-wise product of the Hanning window array, H and each row of the analysis frames matrix, $A[i,:]$. The result is another $k \times N$ array, which is denoted as $\widetilde{A}$, and is given by:

$$\widetilde{A_i} = A_i \odot H \tag{50}$$

We show the effect of a Hanning window applied to an analysis frame:



(a) A section of a violin bowing an $A4$ note, 1024 samples long



(b) The same waveform with a Hanning Window Applied

Window functions are use to account for discontinuities that may arise in the waveform at the edges of each analysis frame. The Fourier Transform assumes that an integer number of signal periods fit within each analysis frame, however this is rarely the case. Transforming a signal with to *spectral leakage* where energy appears to "leak" into adjacent frequency bins. This result is a frequency domain that is not an accurate representation of the equivalent time domain. The Hanning window uses the raised cosine function to taper the signal off to zero at the end-points, thus eliminate discontinuities at the edges while weighting the samples at the center much more heavily. This allows for a much cleaner transform into frequency space.

Before we compute the Fourier Transform to move each analysis frame into frequency-space, we choose to tail-pad each analysis frame with an additional 1024 samples of all zeros. This means that each frame has gone from being 1024 samples, to 2048 samples. Matrix $\widetilde{A}$ has a shape of $k \times 2048$. Doubling the size of each analysis frame allows us the synthetically double the resolution in frequency-space, while retaining the resolution in time-space [32].

4.2.3 Discrete Fourier Transform

With this final change, we perform a *Discrete Fourier Transform* (DFT) to bring each analysis frame from a time domain into a frequency domain [20, 21, 32]. The Discrete Fourier Transform is applied by producing an $N \times N$ *transform matrix*, often noted as $\mathbb{W}$. Let $\omega^k = e^{\frac{-2\pi i}{N}k}$, then the DFT matrix for a time-space signal containing N samples is given by [30, 21, 32]:

$$\mathbb{W} = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \dots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \dots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \dots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \dots & \omega^{(N-1)^2} \end{bmatrix} \tag{51}$$

Each column of the matrix is a complex sinusoidal oscillating with an integer number of periods within the N-sample length window [28, 21]. The DFT is applied by a taking the matrix - product of $\mathbb{W}$ and $\widetilde{A}^T$. The transpose of $\widetilde{A}$ makes each analysis frame into a column vector, which gives the appropriate format for transformation.

$$\text{DFT}\big[\widetilde{A}\big] = \mathbb{W}\widetilde{A}^T \tag{52}$$

Most standard implementations of neural network models use activations, weights, and biases to all be real floating-point numbers [10, 13, 31]. Since the elements of the DFT matrix lie on the complex unit circle, each real-valued analysis frame will be moved into complex space. This means we compute the element-wise product of the transformed signal matrix and it's complex-conjugate matrix. This is the $N \times k$ spectrogram matrix, where each element is a real-values number. Matrix S is defined:

$$S = (\mathbb{W}\widetilde{A}^T) \odot (\mathbb{W}\widetilde{A}^T)^* \tag{53}$$

Where $\mathbb{W}$ is the DFT matrix from Eq. (51) and $\widetilde{A}^T$ is the transpose of the analysis frames matrix from Eq. (50). The asterisks indicate the element-wise complex conjugation. In practice, the matrix product $\mathbb{W}\widetilde{A}^T$ is computed once, and the conjugate is applied to a copy of the array.

The matrix S, is the spectrogram representation of the initial waveform, and has shape $N \times k$ of all real floating-point numbers. We can index matrix S similarly to that of matrix A in Eq. (48):

$$S_{i,j} = S_i[j] = S[i,j] \tag{54}$$

Note that the spectrogram provided to the neural network is representing the power spectrum of the audio file as defined in Eq. (53). All spectrogram figures in this report are from taking the element-wise *natural log* of the power spectrum. This has been done solely for visualization purposes.

Each column of the S matrix is now a single analysis frame that has been moved into a frequency-space representation of itself. There are k columns, just as there were k time-series frames, and N rows since there were N samples per row. Given the discrete nature of digital audio, the frequency-space representation is not a continuous function, but rather a column vector, where the frequency has been assigned to one of N bins, ranging from $-f_s/2$ to $+f_s/2$ [28, 21]. To ensure homogeneous input sizes between all samples, we zero-pad the matrix S with additional columns until $k = 256$, which mimics tail-padding the original signal with zeros. Recall that wave forms were truncated to ensure that $k \leq 256$ analysis frames.

Standard western musical instruments seldom have fundamental frequencies that extend above 6 kHz [20, 32, 34]. This means that when constructing the spectrogram, we will rarely see significant energy present above 12 kHz at any time, and the S matrix will contain mostly zero, or zero-like entries. To condense the size of the matrix, we select only the frequency bins (rows) that correspond to energies between 0 Hz and $12,000$ Hz. This makes the input array smaller, (less than 1/4 the size) and eliminates redundant and non-useful information. The number rows in the S matrix is reduced from N down to N'.

Each spectrogram is now $N' \times k$ and effectively encodes the energy distribution of the waveform as a function of both time and frequency. The spectrogram makes up each sample in the first design matrix X_1, Eq.(44) used in this model. For this classifier, we have chosen $N' = 558$ and $k = 256$. For training, a batch of b samples are concatenated into a single array object. For a batch of b samples of $N' \times k \times 1$ spectrograms, we shape X_1 such that:

$$X_1 = \left\{ S^{(0)}, S^{(1)}, S^{(2)}, ..., S^{(b-1)} \right\} \in \mathbb{R}^{(b \times N' \times k \times 1)} \tag{55}$$

Which is consistent with the shape of the X_1 matrix outlined in Eq. (44). This matrix is presented to the *Convolution* branch of the neural network for processing.

4.3 Time-Space Features

The features described in this section are derived from time-domain representations of each audio sample. For consistency between samples, each waveform is padded or truncated to contain the same number of samples M. The number of samples M is chosen to correspond the same number of samples as needed to make the $N \times k$ spectrogram matrix. In doing this, the spectrogram and time-series features are representative of the same time interval in the file sample. Time space is indexed by $s[i]$ with $i \in [0, 1, 2, 3, ..., M-2, M-1]$

From time space, we use the following 11 features:

- Time Domain Envelope ($\times 5$)
- Zero Crossing Rate
- Temporal Center of Mass
- Auto Correlation Coefficients ($\times 4$)

4.3.1 Time Domain Envelope

The time domain envelope (TDE) is a rough measurement of the energy contained in the time-space of the waveform. If we divide the signal into analysis frames, as in Fig. (17), then the TDE approximates the energy in that frame. A frame with a high TDE indicates a generally large amplitude, and a frame with a low TDE indicates a generally small amplitude, i.e. the signal has not started or is decaying. The small size of each analysis frame makes computing a TDE for each impractical, and creates a very high-dimensional feature vector. This would also cause problems as some waveforms may have an initial attack that differs in time by a few analysis frames. Computing TDE for each frame would therefore introduce temporal bias, as the same waveform delayed by a few milliseconds would generate a slightly different set of predictor values [26].

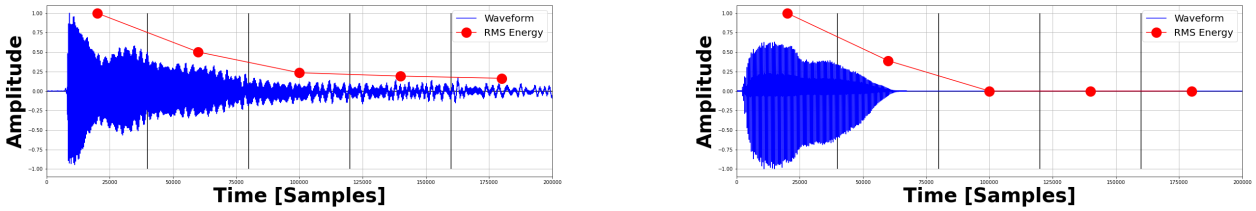
One way to mitigate this is to group analysis frames together and compute the TDE of a collection of frames. This way, almost all waveforms will produce an identical set of features to a short time-delayed version of itself. By choosing a sufficiently large number of TDE values, we can also generate an approximation of than amplitude envelope of the time-series waveform. This comes at a higher computational cost, and a previously mentioned higher dimensional feature vector

We adapt this TDE to **compute the time domain envelope over** 5 **non-overlapping analysis frames**. Since the maximum amplitude of each waveform array has been normalized to ± 1, the TDE represents a consistent measurement of energy in each waveform subset [12]. The TDE is computed as the RMS-Energy of the waveform, s and the j-th TDE value is

given by [20, 26]:

$$\text{TDE}_j\big[s\big] = \sqrt{\frac{1}{Q}\sum_{i=n}^{n+Q} s[i]^2} \tag{56}$$

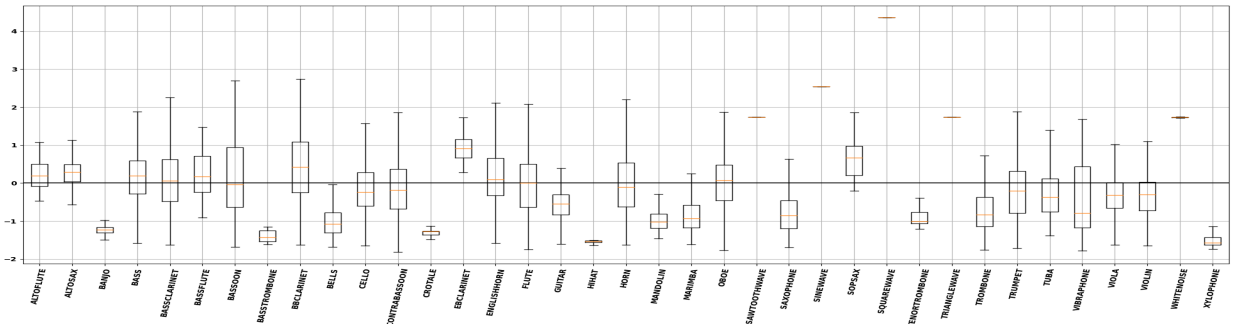
Where Q gives the number of samples in each analysis frame (number of sample sin waveform divided by number of frames), and n gives the index where the j-th frame begins. We provide a graphic representation of how the TDE compares to the amplitude of sections of a waveform in Fig. (19).



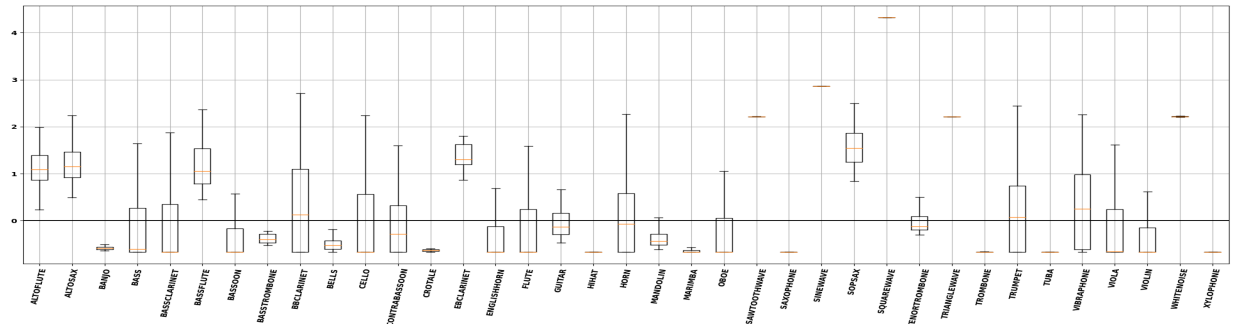
(a) Guitar - B2 (b) Tuba - A2

Figure 19: TDE Envelope values for musical instruments

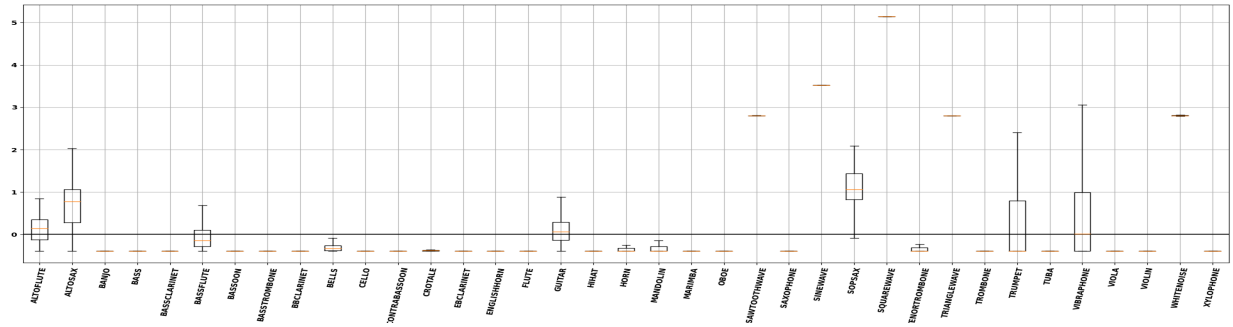
The uniformly sized analysis frames let us create an naive envelope of the waveform, but will allow it to generalize between classes. For example, this allows for a crude approximation of the energy in the attack, decay, sustain, and release portion of the signal [32, 20]. For instruments with heavier attacks, we expect the TDE in the first frame to be comparably large, see Fig. (20a). Instruments with a heavy decay will likely have little energy in the second or higher frames, indicating that the amplitude has substantially died off as in Fig. (20b). Conversely, instruments with longer sustain such as upper woodwinds, vibraphones or strings will contain relatively higher TDE values for the later analysis frames as seen in Fig. (20c). Instruments with little or no waveform envelope, such as the synthesized waveforms and the whitenoise show constant TDE values across each frame as expected.



(a) TDE 1



(b) TDE 2



(c) TDE 3

Figure 20: A comparison of the first three of five time domain envelope values across each class using a box-and-whisker plot

4.3.2 Zero Crossing Rate

The zero crossing rate (ZXR) of a signal or frame is use to measure how many times that a signal crosses it's equilibrium point. This can be computed per total sound wave, per analysis-frame, or per unit time. This feature is most commonly associated with differentiating speech from music, because speech presents a more jagged and often less periodic waveform that musical instruments do [9, 12, 35] .

We adapt this feature to **compute the zero crossing rate for the full waveform**. Signals with a high ZXR can be representative of classes that often have additional noise and signals with a low ZXR, can indicate signals with more stable, periodic behavior. The ZXR for the full waveform s is given by [26, 12]

$$\mathrm{ZXR}\big[s\big] = \frac{1}{2}\sum_{i=1}^{M-1}\Big|\,\mathrm{sign}(s[i]) - \mathrm{sign}(s[i-1])\Big| \tag{57}$$

Where $\mathrm{sign}(x)$ returns $+1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. ZXR also provides a very rough estimate for the average frequency of the waveform. This allows it to allows be useful in discerning classes with generally higher fundamental frequencies, such as upper woodwinds, against classes with generally lower frequencies, such as low brass [12, 34].

This behavior becomes apparent when comparing sets of classes in Fig. (21). For example, consider the ZXR value of a a bass and cellos against that of bells or a soprano saxophone. The bass and cellos have on average much lower fundamental frequencies than that of cellos of soprano saxophones. In the case of this data set, the ZXR provides a predictor which contains clear separations between classes such as crotales and $E\flat$ clarinets.

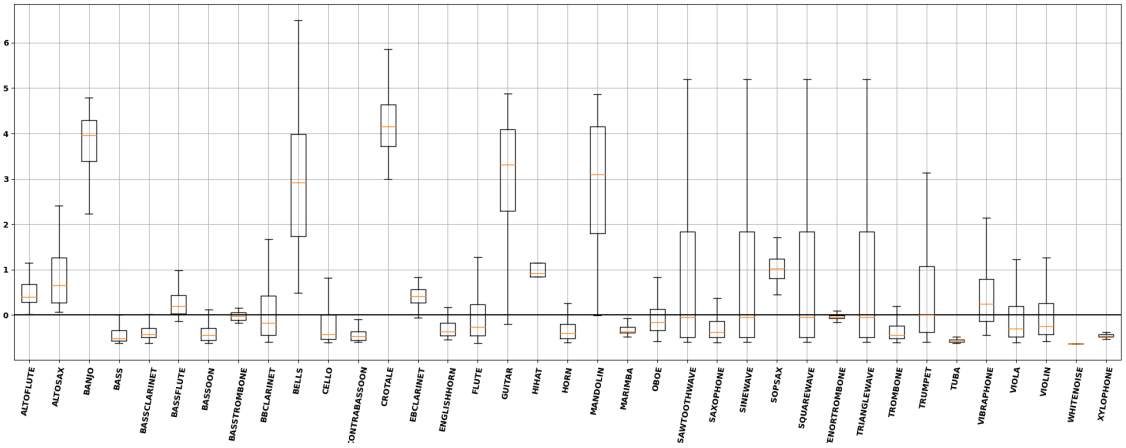


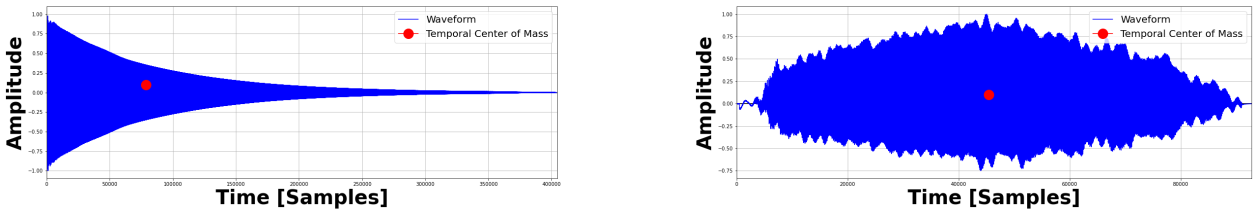
Figure 21: A comparison of the zero-crossing rate for each class using a box-and-whisker plot

4.3.3 Temporal Center of Mass

The temporal center of mass (TCM) of a signal is used to compute roughly where in time the amplitude of the waveform *bunches up*. We compute the element-wise absolute value of the waveform and treat it as a 1-dimensional discrete mass distribution of M samples. The TCM of that waveform is then given:

$$\mathrm{TCM}\big[s\big] = \frac{\sum_{i=0}^{M-1} i\big|s[i]\big|}{\sum_{i=0}^{M-1}\big|s[i]\big|} \tag{58}$$

TCM condenses the idea of the amplitude envelope into a single scalar value. It allows for the quick measurement of where in the time most of the energy of the signal lies. For percussive instruments with short attack and release times, such as bells or xylophones, we expect a very low TCM. Instruments with plucked strings, but longer release times such as mandolins and guitars should have similarly low values. See Fig. (22) for examples.



(a) Vibraphone at $A5$ shows a lower TCM

(b) Cello at $F4$ shows a central TCM

Figure 22: TCM values for musical instruments

Instruments in the woodwind and brass family generally have longer sustain and release times, giving them a slightly higher center of mass. Strings and undampened percussion have notoriously long sustain and release times giving them much higher TCM values [20, 34]. Finally, the synthetic wave forms have no characteristic envelope shape which gives a centrally located TCM.

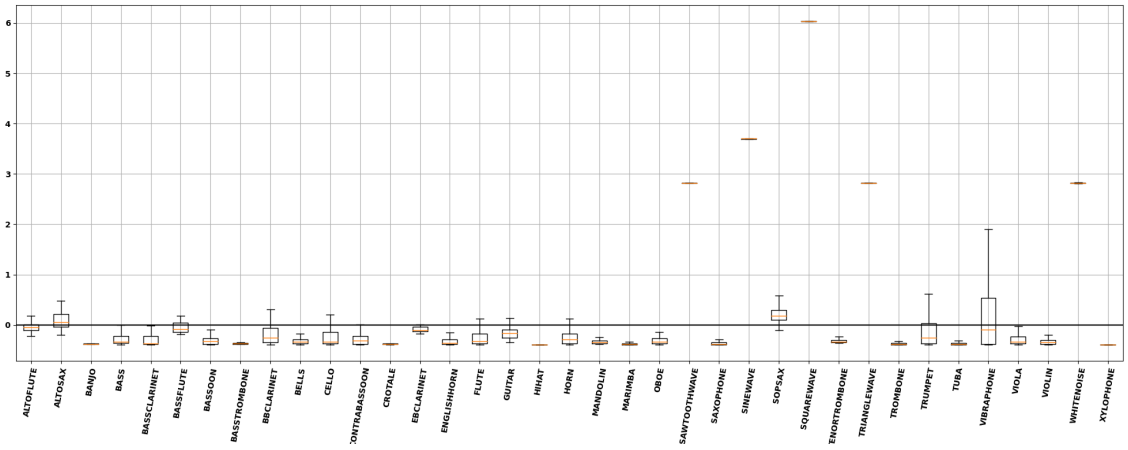


Figure 23: A comparison of the temporal center mass for each class using a box-and-whisker plot

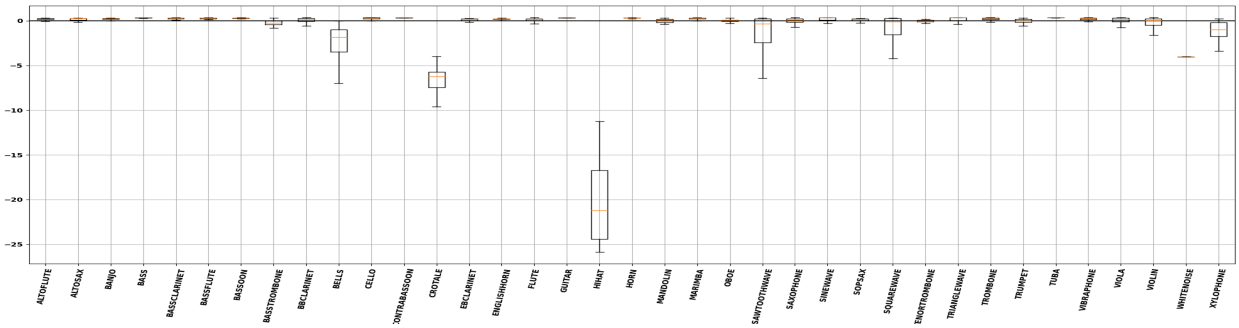
4.3.4 Auto Correlation Coefficients

Auto correlation coefficients (ACC) are rough estimates of the signal spectral distribution. They are computed by multiplying a signal with a time-expedited variant of itself, and then normalized to be between 0 and 1 We can compute any number of ACC's and their value

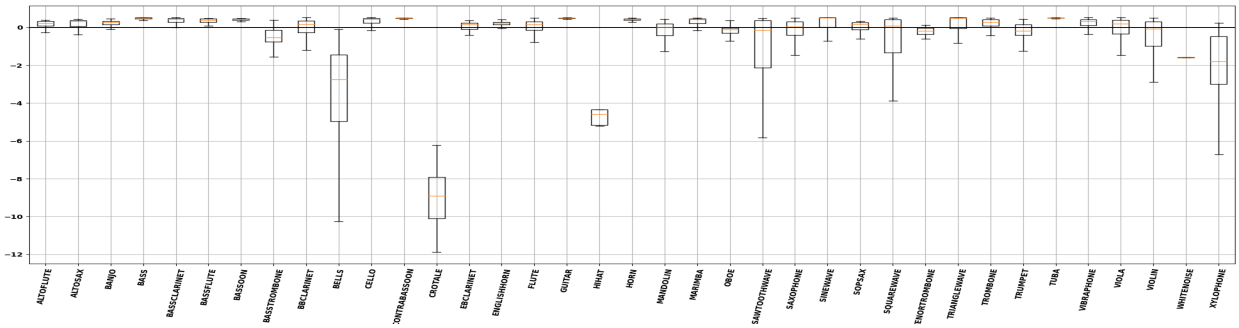
changed depending on the index chosen. It is common to use the first K ACC's [26]. For a full waveform signal s, with M samples, the k-th ACC (indexed from 1 to K) is given by:

$$\mathrm{ACC}_k\big[s\big] = \frac{\sum_{i=0}^{M-k-1} s[i]s[i+k]}{\sqrt{\sum_{i=0}^{M-k-1} s^2[i]}\sqrt{\sum_{i=0}^{M-k-1} s^2[i+k]}} \tag{59}$$

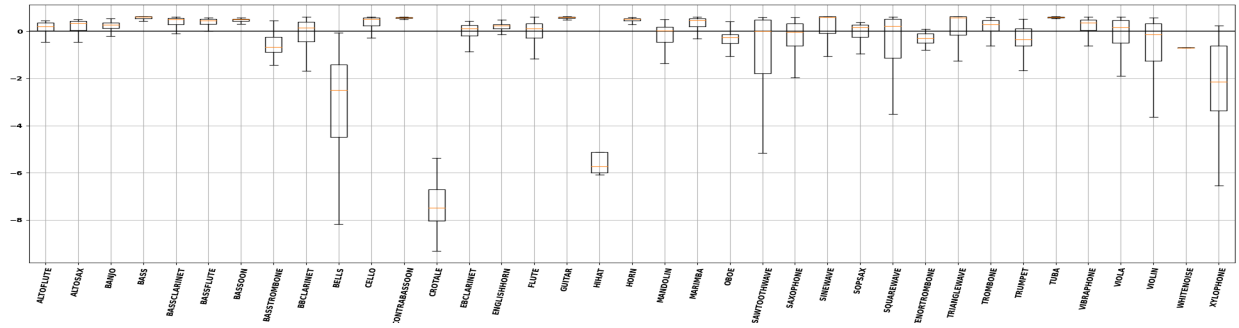
Physically, the numerator of the ACC representing computing the dot product of the signal, $s[i]$ with a time-hastened version of itself $s[i+k]$, and the denominator provides a normalization for the value. Dotting the signal and the time delay allows us to introduce a synthetic phase shift and compare the resulting relationship. If we chose k to be equal, or similar to the number of samples that make-up a period or half-period of the waveform, then $s[i] \approx s[i+k]$ and then $\mathrm{ACC}_k \to 1$. Alternatively, for many other values of k, frequent multiplication and summation of samples that are approximately zero result in $\mathrm{ACC}_k << 1$. This make auto-correlation coefficients extremely useful for detecting periodicity in time-space [26].



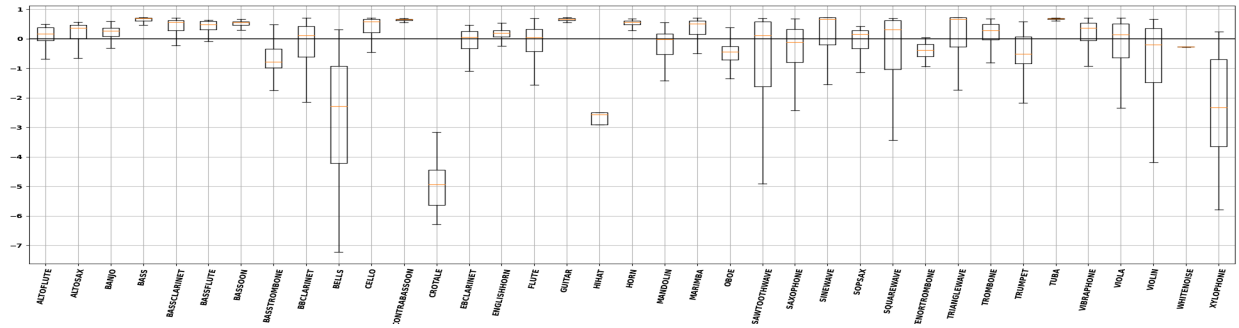
(a) ACC $k = 1$



(b) ACC $k = 2$



(c) ACC $k = 3$



(d) ACC $k = 4$

Figure 24: A comparison of the first four auto correlation coefficients in each class using box-and-whisker plots

4.4 Frequency-Space Features

The features described in this section are derived from the frequency-domain representations of each audio sample. For each feature, we detail the physical significance and provide a visualization in feature-space.

From frequency space, we use the following 13 features:

- Mel Frequency Cepstral Coefficients ($\times 12$)
- Frequency Center of Mass

4.4.1 Mel Filter Bank Energies

Mel Filter Bank Energies (MFBE's) are not used directly as features, but are used in computing Mel Frequency Cepstrum Coefficient (MFCC's) so we describe them here. Mel filter banks are divisions of the frequency spectrum of a signal into R overlapping bins [25, 26]. These filter banks allows us to group sounds based on their energy distribution in frequency space. Each filter is triangularly shaped, covering a certain band in frequency space, and zero elsewhere. This way, when computing the the dot product of any filter with frequency space, we get an approximation of energy in that filter bank [25, 26].

Rather than producing filter banks based on the linear Hertz scale, the frequency axis of the signal is transformed into units of *Mels*, which is used to account for the non-linearity in human pitch perception [26, 9, 20]. Filter banks are produced to be evenly spaced on the Mel scale, and then transformed back into the Hertz scale. This has the effect of producing triangular filter banks with grow in width as the frequency increases. The Hertz to Mel and Mel to Hertz transforms are given [26, 9]:

$$M_f[h] = 2595 \log_{10}\left(1 + \frac{h}{700}\right) \tag{60}$$

$$H_f[m] = 700\left(10^{\left(\frac{m}{2595}\right)} - 1\right) \tag{61}$$

Where M_f is the frequency in units of Mels, given $[h]$, a frequency in Hertz, and H_f is the frequency in Hertz given $[m]$ a frequency in Mels.

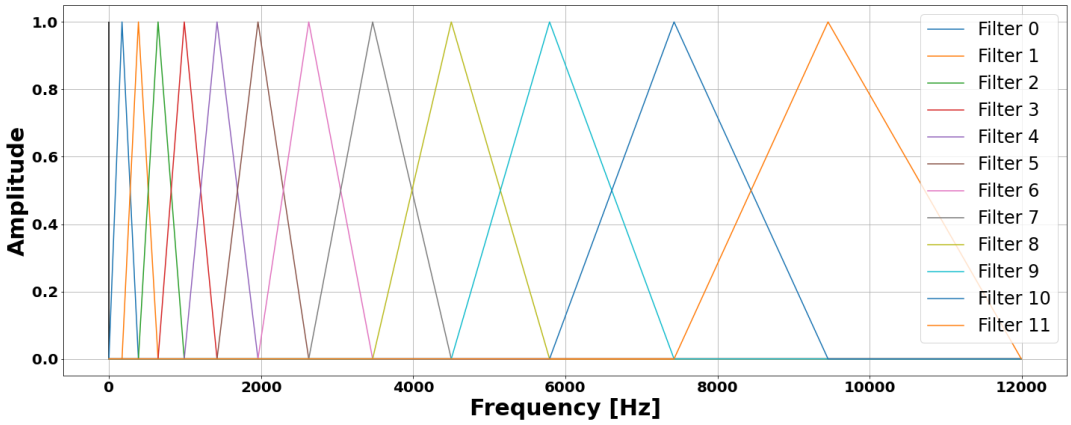


Figure 25: Mel Filter Banks shown in frequency space with units of Hertz

Each of the R filters is created to be N' samples long, to match the width of the cropped frequency space in the spectrogram, Eq. (53). When applied to an analysis frame in the frequency spectrum, the dot-product between the filter and the spectrum gives an approximation of the energy in that filter bank. Each filter is concatenated into a matrix M of shape $R \times N'$, where each row is one filter. We apply the Mel Filter banks to the spectrogram to create matrix B:

$$B = (MS)^T \tag{62}$$

Matrix B has shape $k \times R$.

The matrix product in Eq. (62) allows that $B_{i,j}$ is the dot product between the i-th analysis frame and the j-th filter-bank. Finally, we compute the average energy across all k frames, into array $\widetilde{B}$ with shape $1 \times R$. For this project, we have chosen to use $R = 12$ filter-banks, which are all are used to compute the MFCC's in the next section.

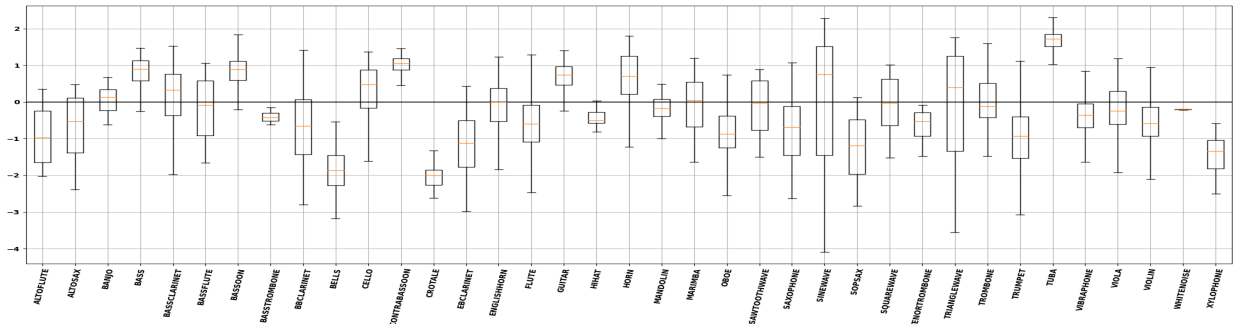
4.4.2 Mel Frequency Cepstral Coeffecients

Ceptstral coefficients are the result of computing the inverse discrete Fourier transform (IDFT) of the logarithm of the frequency Spectrum [26, 25]. Mel Frequency Cepstral Coefficients (MFCC's) are the most commonly used ceptral coefficients and appear often in digital signal processing. The c-th coefficient is given by:

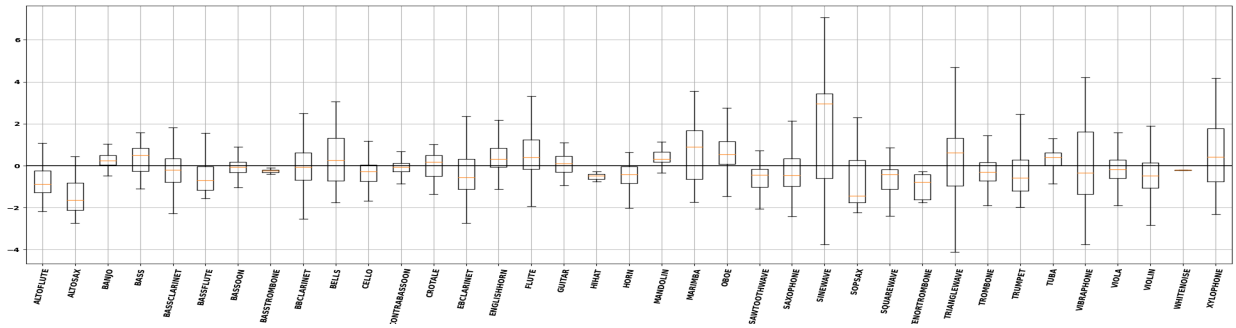
$$\text{MFCC}[c] = \sqrt{\frac{2}{R}} \sum_{i=1}^{R} \log\left(\widetilde{B[i]}\right) \cos\left(\frac{c(i-\frac{1}{2})\pi}{R[i]}\right) \tag{63}$$

Where $\widetilde{B}$ is the column average of the Mel filter bank energies computed in Eq. (62), and R is the number of filter banks used.

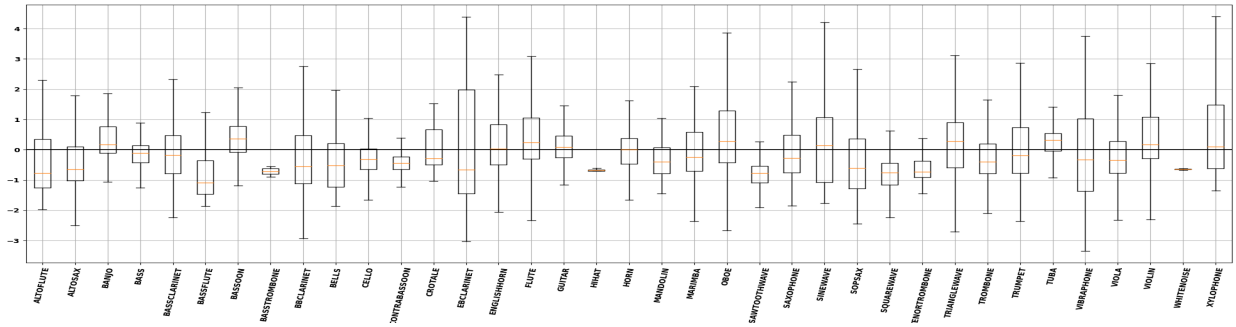
Physically, MFCC's are a b inverse transform of a transform. This allows us to investigate the periodicity of the frequency spectrum, which highlights phenomena such as overtones or echoes in the signal [32]. Cepstrum coefficients are commonly used for speech identification and are very prolific in sound recognition tasks [26, 25, 12]. We show a feature-space representation of MFCC's for a selection of coefficients in Fig. (26).



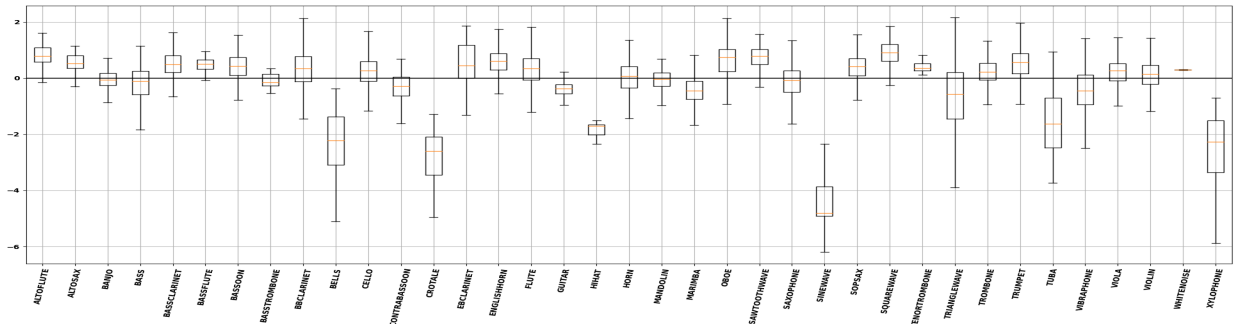
(a) MFCC $R[1]$



(b) MFCC $R[4]$



(c) MFCC $R[8]$



(d) MFCC $R[12]$

Figure 26: A comparison of 4 Mel Frequency Cepstral Coefficients, 1, 4, 8 and 12 for each class using box-and-whisker plots

4.4.3 Frequency Center of Mass

The frequency center-of-mass (FCM) for an audio file provides a representation of how overtones and energy is distributed in the signal's frequency domain. As with the temporal center-of-mass, we treat each column of the S matrix as it's own one-dimensional mass distribution, and compute the center of mass of each column. This encodes the FCM for a single analysis frame (in frequency-space) in the waveform. For an frequency-space analysis frame, $s^{(i)}$, the FCM is given by:

$$\text{FCM}_i\left[s^{(i)}\right] = \frac{\sum_{j=0}^{N'-1} j s^{(i)}[j]}{\sum_{j=0}^{N'-1} s^{(i)}[j]} \tag{64}$$

We compute the FCM for each of the k' frames, and then average the results. We use the average FCM across k' frames to compute the FCM feature:

$$FCM = \frac{1}{k'} \sum_{i=0}^{k'} \text{FCM}_i\left[s^{(i)}\right] \tag{65}$$

The average FCM gives a strong approximation of the instrument or signal source's range. For example, a flute or violin will have a considerably high FCM value, even in their lower registers. Similarly, basses or tubas will have considerably low FCM values. Given that the standard frequency range of some musical instruments is fixed, it is expected that for any particular instrument, the FCM will consistently have low variability [20, 34].

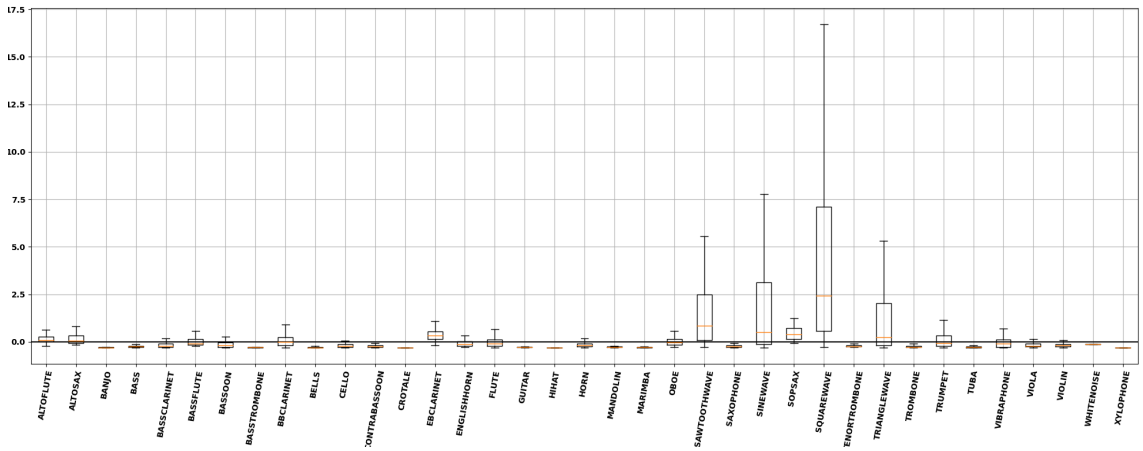


Figure 27: A comparison of the frequency center of mass for each class using box-and-whisker plots

5 Evaluating Model Performance

Before making predictions on unlabeled data such as the Chaotic Synthesizers, we must confirm that our model performs reasonably well on data that it has never interacted with. The most common practice is to divide a full data set into a subset of *training* samples, and *testing* samples [2]. As the names imply, the training subset is used to fit the model, and we use the labeled testing data set to evaluate how well the model has trained. The exact ratio of sizes between these subsets varies depending on the task [4, 3, 17], however we choose to generally use around 90% training samples and 10% testing samples. If the model performs well within it's training subset, but poorly within the testing subset, this can indicate a model that is *overfitted*, and will not perform well on other unseen samples [2].

5.1 K-Folds Cross Validation

We can expand on the idea of a train/test split with a resampling method called *K-Folds Cross Validation* (Also called X-validation) [2, 8]. Suppose we have machine learning model F^*, with a set of trainable parameters Θ. We also have a data set X and a set of appropriately labeled targets, Y, each of which contain N samples. For K-folds X-val, we divide the full data set into K subsets (called *folds*), each with size N/K:

$$X, Y \rightarrow \left\{ \left(X^{(0)}, Y^{(0)}\right), \left(X^{(1)}, Y^{(1)}\right), ..., \left(X^{(K-1)}, Y^{(K-1)}\right) \right\} \tag{66}$$

For each iteration, $k \in [0, 1, 2, ..., K-1]$ we reserve a single subset of data $\left(X^{(k)}, Y^{(k)}\right)$ to use as a testing subset. We use the remaining $K-1$ subsets as training data. We fit the model with the data and labels to produce a model $F^{*(k)}$ which as set of parameters Θ^k. We return the the test set, and run predictions on this unseen data subset. We compare those predictions to the corresponding set of labels $Y^{(k)}$ and evaluate the result of any selected performance metrics.

The output of cross validation is K models that are all trained and evaluated on overlapping subsets of the full data. Each model F^*, shows a possible outcome of training the network given a subset of data samples, and a particular set of initial parameters. We can compare the performance of each model and test if they produce similar outcomes, which indicates that the model will be able to consistently generalize to new, unseen samples.

Algorithm 7 A K-Fold Cross Validation program.

Require: Untrained Network or related learning algorithm, F^*
Require: A full labeled data set (X, Y) of N samples
Require: Number of splits in Cross validation, K
Require: Performance metric function(s), P
  Divide Data into K non-overlapping subsets x_i, each with roughly N/K samples
  $X, Y \rightarrow \left\{ \left(X^{(0)}, Y^{(0)}\right), \left(X^{(1)}, Y^{(1)}\right), ..., \left(X^{(K-1)}, Y^{(K-1)}\right) \right\}$
  Performance History $\leftarrow \{\}$
  for $k = 0, 1, 2, 3, ..., K-2, K-1$ **do**
    Reset all parameters in F^* to a random "untrained" state
    Set aside testing data subset
    $X_{test} \leftarrow X^{(k)}$
    $Y_{test} \leftarrow Y^{(k)}$
    Concatenate the remaining subsets into training data set
    $X_{train} \leftarrow X^{(i \neq k)}$
    $Y_{train} \leftarrow Y^{(i \neq k)}$
    Train the model, F^* with the X_{train} and Y_{train} subset.
    Evaluate the trained model with the X_{test} and Y_{test} data set, and compute value of metric function(s) P
    Store Performance P in Performance History array
  end for
  Compare performance results, and adjust model or parameters as needed, and repeat if desired.
  Run additional analysis on performance metrics.

Cross Validation is particularly useful in models such as neural networks because they suffer from the phenomenon of *high-variance*, which means that small changes in initial conditions can drastically change the outcome of the model [8]. Cross validation allows us to control these initial conditions by training the model on similar, but non-identical subsets of data, along with a slightly different set of initial parameters. This repetition allows us to explore the model's behavior over a range of possible initial conditions and validation sets, to ensure that the network is functions as expected over multiple trials. This also eliminates the fear of "unique" cases where the model happens to perform exceptionally well or exceptionally poorly given a random set of initial parameters [2].

5.2 Performance Metrics

In the case of the multi-category classifier, it is important that we choose the appropriate performance metrics to confirm that the network is completing it's assigned as as expected [2]. While the neural network itself uses the cost function as it's sole objective to optimize, we also require a set of more human-readable functions. For example, an average loss score of 1.075 over a given subset of previously unseen samples provides us with no real information

as to how well the network is performing at it's designed task. In this section, we introduce a set of functions and metrics that allow for a more tangible interpretation of the model's performance. To evaluate any performance metric, we require a set of samples with ground truth labels, y, and a model's prediction for those labels, y^* [4, 8].

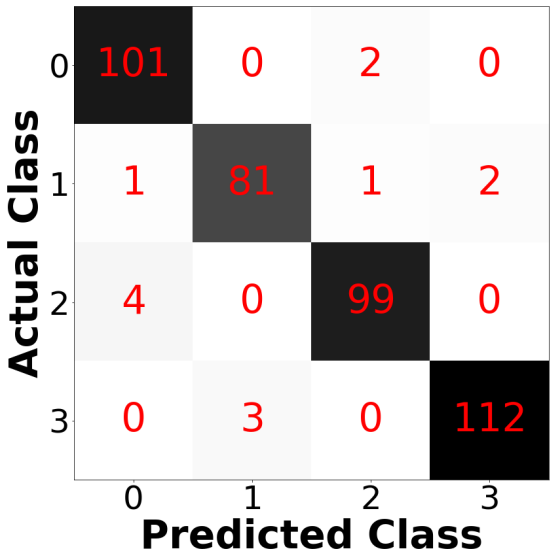
5.2.1 Confusion Matrix

The *standard confusion matrix* (also called a confusion table) is a very quick, often graphical model that can be used to show how a classifier model performs over a subset of predictions. The general idea of this object is count the number of times class i is predicted to be in class j, and vice-versa [2]. If we see that these classes are being repeatedly *confused* in the model's prediction process, then we can modify the model or features to account for it, or explore why this happens.

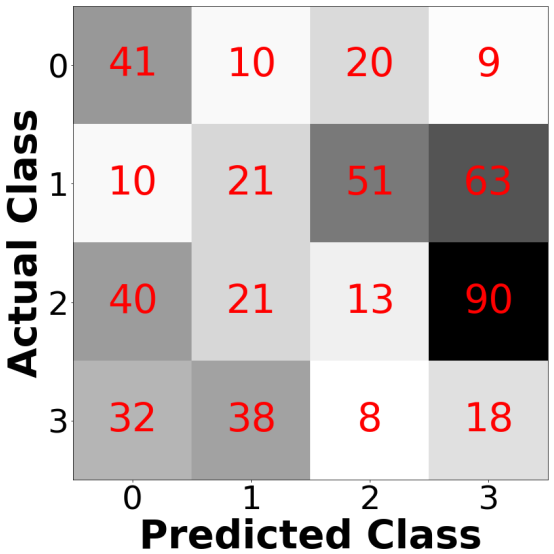
For a k-classes classifier, a confusion matrix will have shape $k \times k$, and every element is a non-negative integer. Each row represents the "ground truth" or labeled class, and each column represents a predicted class. Thus for a confusion matrix, C, we can say that:

$$C_{i,j} = \text{Number of samples that}$$
$$\text{belong to class } i, \text{ and were predicted to be in class } j$$

Thus, indexes where $i = j$ represents a correct prediction, and $i \neq k$ indicates an incorrect prediction. A confusion matrix with relatively large values in main diagonal indicates a model that predicts correct labels [2]. Below we present some synthetic confusion matrices, which combine the counts in each index and a corresponding color map.



(a) Diagonal dominance indicates a stronger classifier



(b) No diagonal dominance indicates a weaker classifier

Figure 28: Example Confusion Matrices for 4-categories classifier

In addition to the standard confusion matrix, we can also weight entries by prediction score. Rather than adding +1 to each entry $C_{i,j}$ where appropriate, we instead add the *probability* value, which is bounded $[0, 1]$ (See softmax activation function in section 2.4.2)). This allows us to bake in the predictive *confidence* into the confusion matrix. For example, two correct predictions with a score 0.2 are weighted equally to a single incorrect prediction of 0.4. This is useful for identifying a classifier that makes correct predictions, but only by smaller threshold. This can almost be through of as assigning "partial credit" to each prediction.

However, it is more useful to weight the confusion matrix according to class occurrence. For example, if we had a 2×2 matrix with entries 100 and 20 in the main diagonal, it seems that class A has a much higher prediction success than class B. However, if there are 200 elements in class A, and only 21 in class B, then we see that class A has a 50% classification rate, while class B has a 95% classification rate. Consider the list of instrument classes in Fig.(7) and notice how all instruments are not equally represented. This creates a training bias such that the model is more heavily trained on a particular class compared to others. This may also show when making predictions, where the classifier is more likely to predict samples that it has been trained on more [8, 13]. To Combat this, we typically normalize each confusion matrix by the number of samples in each "actual" class. Put simply, we divide each element in a given row by the *sum* of that row.

The confusion matrix can be a cumbersome, so often it is useful to express the performance in terms of more concise quantities. For this we use *accuracy score*, *precision score*, *F1-score*, and *recall score*. These metrics are standard in the field of machine-learning classification and are often useful for identifying different strengths and weaknesses in each model [2, 8].

5.2.2 Accuracy Score

Accuracy score, while not commonly used is the most intuitive of all of the performance metrics. It is the ratio of correct predictions of he total number of predictions. While immediately useful for equally sampled binary classification problems, it looses meaning as the number of classes increases, and epecially if the number of samples per class is not consistent [2].

For example, we could develop a model that uses a few predictors from every human currently alive and determine which of them has walked on the moon. There are roughly 7 billion humans alive, four of which have been on out moon. Simply guess *no* for every human, our model would have more than a 99.9999% accuracy. However, we would argue that this model does not perform well seeing as it has a 0% recall score, and and a 0% precision score.

Accuracy score of model is defined:

$$\text{Accuracy} = \frac{TP + FN}{TP + FP + FN + FP} \tag{67}$$

For a multi-class problem, present an accuracy for each individual class, called *micro-accuracy* [8]. The accuracy for a class j is the correct predictions related to the class divided by the total predictions and total counts of the class. We express this mathematically by using sums over each row, each column, and the main diagonal as such:

$$\text{Accuracy}_j = \frac{\sum_i C_{i,i}}{\sum_i C_{i,i} + \sum_{i \neq j} C_{i,j} + \sum_{i \neq j} C_{j,i}} \tag{68}$$

Where all sum indexes goes from 0 to $k-1$.

We can also present *global accuracy* or *macro accuracy* by computing the acurracy over the full confusion matrix. This would then by the ratio of all correct predictions (the main diagonal) to all predictions (the sum of the matrix). We can express this as:

$$\text{Accuracy} = \frac{\sum_i C_{i,i}}{\sum_i \sum_j C_{i,j}} \tag{69}$$

5.2.3 Precision Score

Precision score (also called *specificity* or *positive predictive value*) is the ratio of chosen elements to all relevant elements. This bounds precision to the range $(0, 1)$, with a higher value more desirable. For a classifier with $k = 2$ unique classes, we define the precision score of a model as:

$$\text{Prec} = \frac{TP}{TP + FP} \tag{70}$$

Where TP is the number of *true-positive*, and FP is the number of false-positive predictions. For a k-classes confusion matrix, C, the precision score of a class j, is given by the entry $C_{j,j}$ divided by the sum of row j:

$$\text{Prec}_j = \frac{C_{j,j}}{\sum_{i=0}^{k-1} C_{j,i}} \tag{71}$$

In the case of a multi-class classifier, the precision score represents a *one-vs-all* measurement. This means that for any class j, the sample belongs to class j or it does not. All other classes, $i \neq j$ are temporarily considered to be a single aggregated class. The quantity $TP + FP$, or the sum over the confusion matrix row is a measurement of the total number of items that are predicted to be in the given class. Therefore the precision metric answers the question: *"How many selected items are relevant to the problem?"* [2, 8].

5.2.4 Recall Score

Recall score (also called *sensitivity* or *true positive rate*) also offers a more concise performance metric than a confusion metric. For a classifier with $k = 2$ unique classes, we define the recall score of a model as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{72}$$

Where TP is the number of *true-positive*, and FN is the number of *false-negative* predictions. For a k-classes confusion matrix, C, the precision score of a class j is given by the entry $C_{j,j}$ divided by the sum of column j:

$$\text{Recall}_j = \frac{C_{j,j}}{\sum_{i=0}^{k-1} C_{i,j}} \tag{73}$$

In the case of a multi-class classifier, the recall score also represents a *one-vs-all* measurement. This means that for any class j, the sample belongs to class j or it does not. All other classes, $i \neq j$ are again temporarily considered to be a single aggregated class. The quantity $TP+FN$, or the sum over the confusion matrix coumn is a measurement of the total number of items that are actually in the given class. Therefore the recall metric answers the question: *"How many relevant items to the problem are selected?"* [2, 8].

5.2.5 F1-Score

F1-score (also called F-score) is the harmonic mean of the precision and recall scores and it also bounded on $[0, 1]$ with a higher score being favorable [2]. Often, the two metrics can be through of as somewhat *exclusive* to each other - the higher one, the lower the other. Compare this with the idea that the more *sensitive* the model is, the less *specific* it is, and vice-versa. This phenomenon is called the *precision-recall tradeoff* [2, 8]. We compute the $F1$ score for a class j as:

$$\text{F1}_j = 2 \times \frac{\text{Prec}_j + \text{Recall}_j}{\text{Prec}_j \times \text{Recall}_j} \tag{74}$$

The $F1$ favors models with both a high precision and high recall score. Some models allow for the ability to adjust the threshold of the decision function. This means that by changing a few parameters that are external to the classifier itself, we can change how sensitive or specific a model is when making it's class decision. In the case of this multi-category classifier, this is somewhat like modifying the generated decision boundaries as to change the outputted probability distribution. The $F1$ score to find the set of hyper-parameters that allow for the classifier to produce both a high precision *and* recall score.

5.3 Tracking Metrics over a Period of Training

A period of training is characterized by fitting the parameters Θ of a model F^* to a set of data X and corresponding labels Y [4, 32]. This is done by passing subsets of data, called *mini-batches*, into the neural network for training. The average cost for the mini-batch is then used to compute the gradient vector $\nabla_\Theta J$, and subsequently update the model according to the optimizer chosen [2, 4]. Recall we implement an ADAM optimizer as outlined in section (2.5.4).

We execute training computations in batches to reduce the amount of memory required for the operation. Pushing a full data set through a model at one time would require more

RAM than is usually available so we use repeating, non-overlapping subsets of data called mini-batches to avoid memory exhaustion errors. Each mini-batch used equates to one step in the optimizer update rule in Eq. (37) [4]. Thus, a single pass over a full data set allows for multiple iterations of the optimizer to reduce the value of the cost function. This can be combined with multiple passes over the full data set, called *epochs* [8, 13].

Large mini-batches require lots of RAM, but prevent the model from being over-fit to to any one sample, or class of samples. Small mini-batches require less RAM, but often may bias the optimizer to over-fit the given samples, and make optimization unstable [2, 8]. To ensure that the model is optimizing properly, we can track how the metrics behave over a training period. This allows us to monitor the *rate of convergence* of the model. If the cost function is dropping to quickly or slowly, this may indicate that the optimizer learning rate is too high or too low. This may lead to an over-fit or under-fit model, or indicate an inappropriate set of features are being used [2, 4]. In Some cases, it can be used to initiate *early-stopping*, which halts training when a set of conditions are met.

Given the high dimensionality of this model's parameter space,the large volume of sound files, and the large amount of RAM required to store the spectrogram matrix and the feature vector for each sample, we do not directly load the full data set into memory at once. Instead, we produce a large subset of the full data set of 256 samples which we dub a *mega-batch*. From this mega-batch, we produce the two design matrices required for input, and load the corresponding labels. We then use a subset of *this* group as the *mini-batches* for training. Each mini-batch contains 32 samples. Once all mini-batches are fit, we discard the mega-batch design matrices and repeat for the next 256 samples.

With each training step, we have recorded the precision, recall, and loss scores. These values are computed from the forward pass of the training data before the gradient is computed, and parameters updated, i.e. the program has not seen these particular sample system. Our program stores past metric and loss score histories locally in a *training-history* file, which we can examine after the program completes. Below, we visualize the evolution of each score as training progresses. These plots are from a model that was trained separately on the full data set after cross validation was performed.

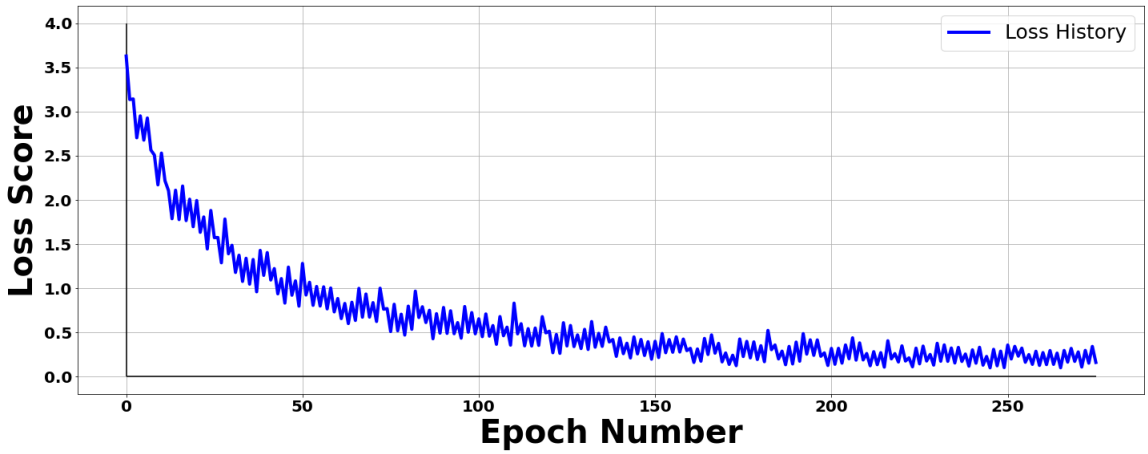


Figure 29: The loss function score decreases with each training step, indicating that optimization is performing correctly

similarly, we can visualize the precision score and recall score at each training step.

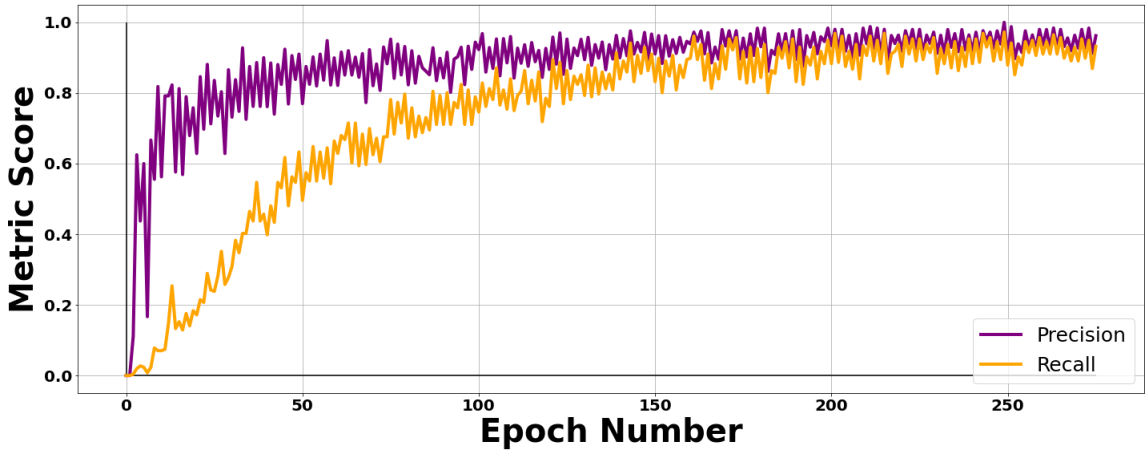


Figure 30: The precision score and recall score generally increase with each training step, indicating that the model is getting both more sensitive and more specific as training progresses

Notice how loss, precision and recall show a pattern of increasing and decreasing every two steps. This is because we extract a mega-batch of 256 samples, and perform two epochs of training on the batch. The first step represents the first time the model has seen the data, thus produces a higher cost function value. In the second step, the model has already seen the samples in the batch, and it produces a slightly lower cost value that the first step. The batch is discarded and a new subset of samples is drawn so that the process repeats. With

two epochs per mega-batch, and two passes over the full data set (permuted in the middle), the model is effectively trained on the full data set a total of four times.

6 Experimental Results

6.1 Executing Cross Validation

To formally produce classification predictions on the chaotic synthesizer waveforms, we must use all of the neural network, physics, and statistical principles outlined in the previous sections. Using the described features in section (4) and appropriate architecture section (2.6), We run a $K = 10$ folds cross validation program as in Alg.(7). For each of the 10 models, we produce a standard confusion matrix, and compute the (i) accuracy score, (ii) precision score, (iii) recall score, and (iv) F1 score. In Fig. (31) we show how the metrics compare in each the 10 models by averaging the scores across the 37 classes.

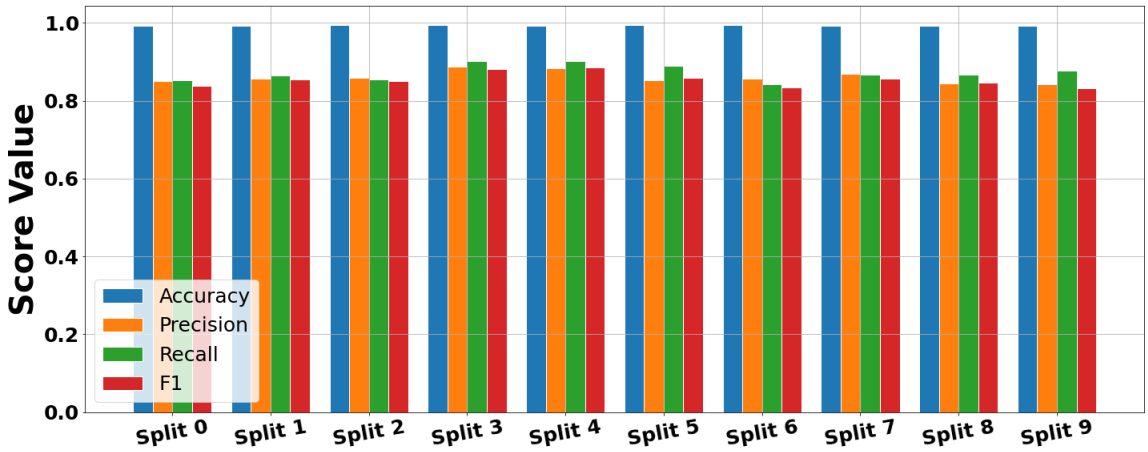
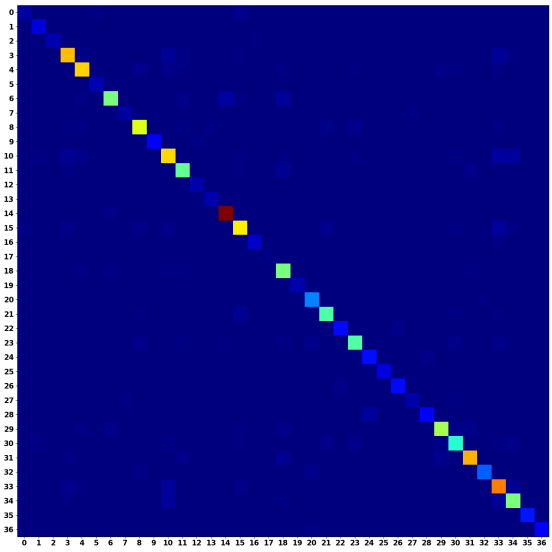
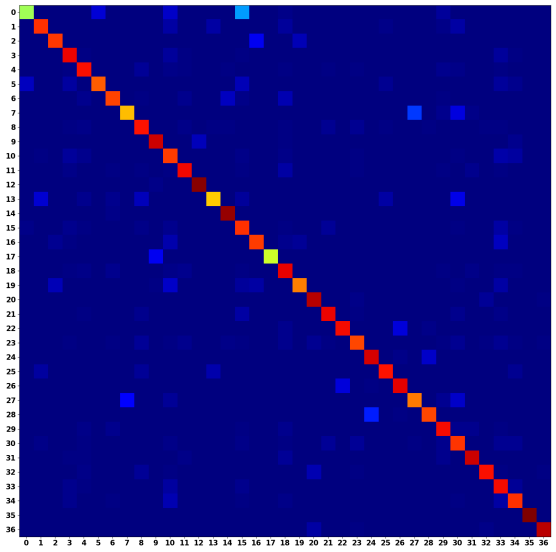


Figure 31: Performance metrics for the multimodal networks across 10 models, scores are averaged over 37 classes

Additionally, we present the average standard confusion matrix, and the hits-weighted confusion matrix in Fig. (32).



(a) Standard confusion matrix for hybrid network



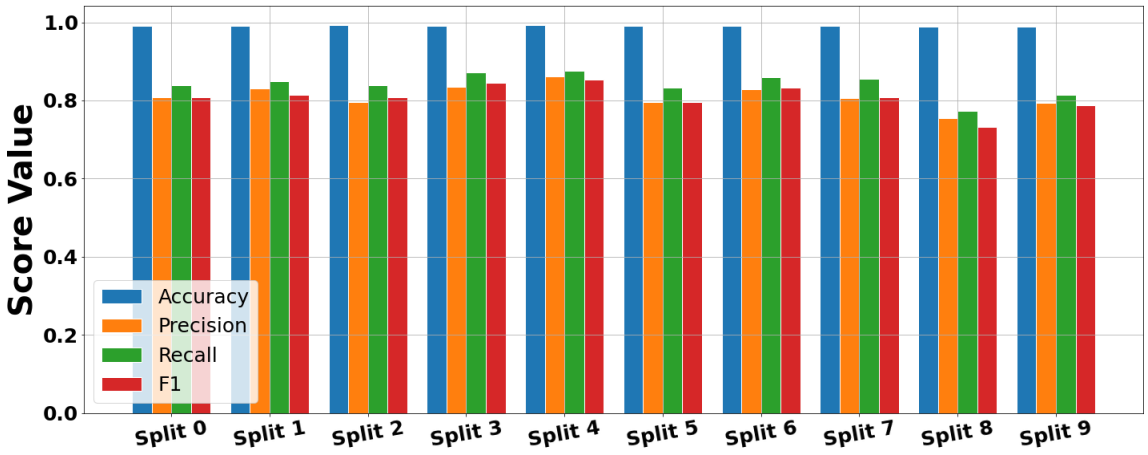
(b) Occurrence weighted confusion matrix for hybrid network

Figure 32: Confusion matrices, each is averaged over 10 folds of cross validation. Recall that each integer represents a class labeled, as given in Tab. (7)

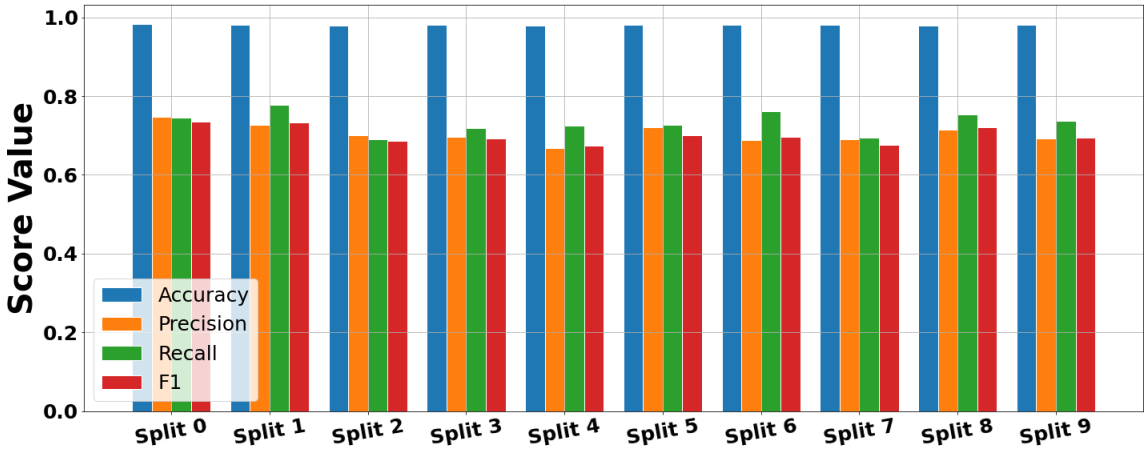
We show the metric values of all splits to show the consistency in performance across all subsets. Notice how each split shows reasonable performance as given by the precision and recall score consistently above 80%. This indicates that for this data set, architecture, and feature selection, our model can accurately classify waveforms and generalize appropriately. This step is critical in ensuring that there are no major out-liers in the classification results, and that any set of initial conditions can still allow the model to reach an acceptable set of parameters and decision boundaries. The results of cross validation indicate a strong classier that can be implemented to make predictions on other waveforms.

6.2 Comparing Results between Architectures

In addition to considering our multimodal network architecture performance, we have also run the identical $K = 10$ folds cross validation program on two uni-modal variants of the model. The first variant contains only the Convolutional branch of the network in Fig. (6), meaning that we feed the activations from the last dense layer in the left column directly to the output layer. Similarly, the second variant contains only the perceptron branch of the network in Fig. (6), meaning that we feed the activations from the last dense later in the right column directly to the output layer. For consistency, the same full data set (around $18,000$ samples) was using in the cross validation program. Comparing the performance of the each model gives us an insight as to the predictive contribution of each input branch



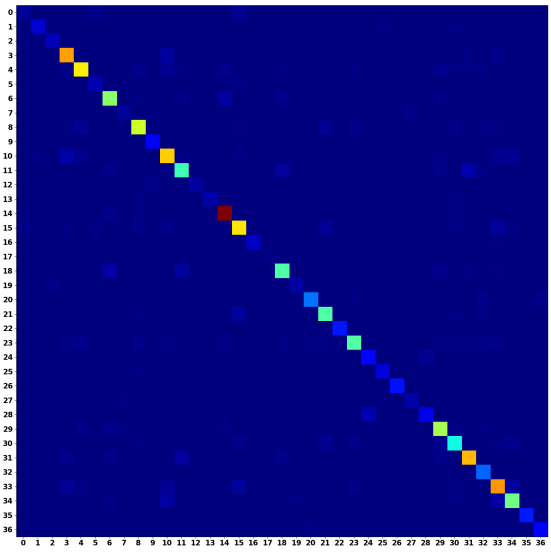
(a) Convolutional Unimodal Architecture, see left side of Fig. (6)



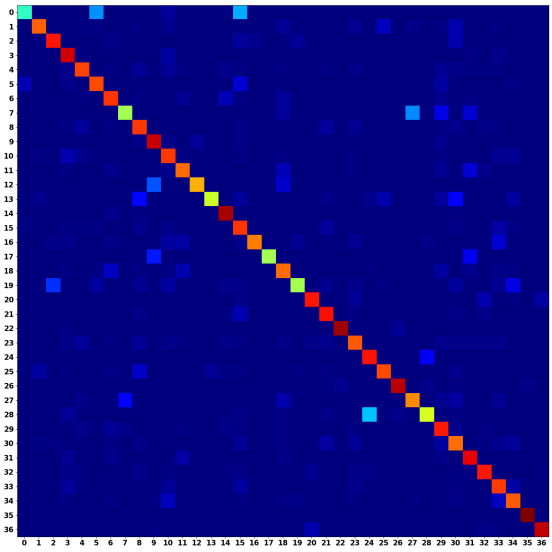
(b) Perceptron Unimodal Architecture, see right side of Fig. (6)

Figure 33

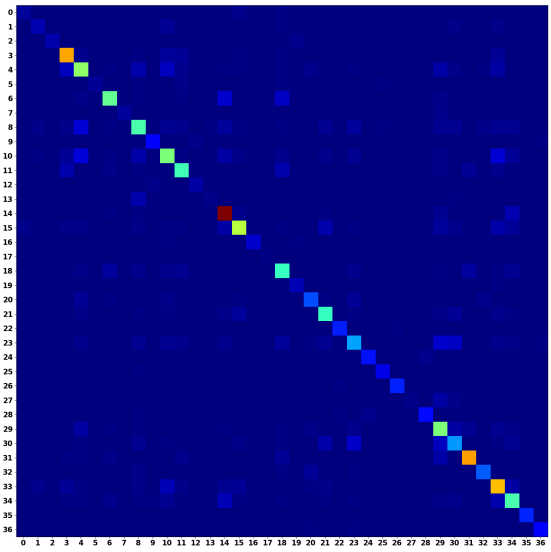
Compare these unimodal model performances with that of the multimodal network results in Fig. (31). Notice how the multimodal network consistently produce superior classification results. Similarly, we present the standard and weighted confusion matrices for both unimodal networks. Compare the results with Fig. (32) and see how the confusion matrix indicates a stronger classifier.



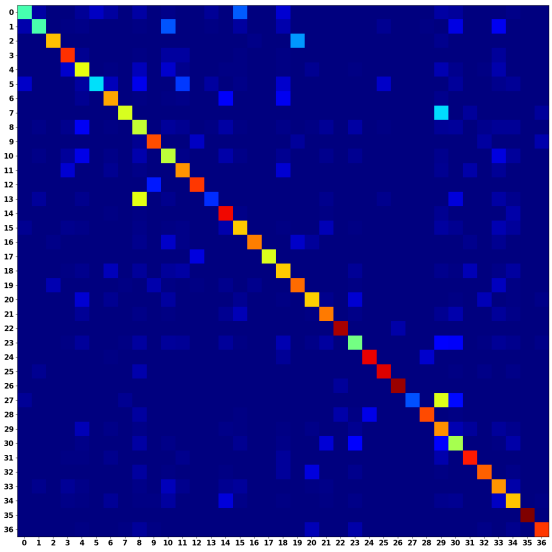
(a) Standard confusion matrix for the convolution branch



(b) Occurrence weighted confusion matrix for the convolution branch



(c) Standard confusion matrix for the perceptron Branch



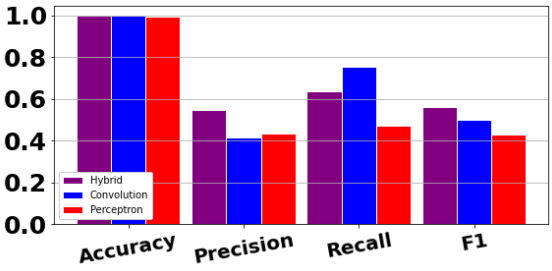
(d) Occurrence weighted confusion matrix the for perceptron branch

Figure 34: Performance metrics for the unimodal networks across 10 models, scores are averaged over 37 classes

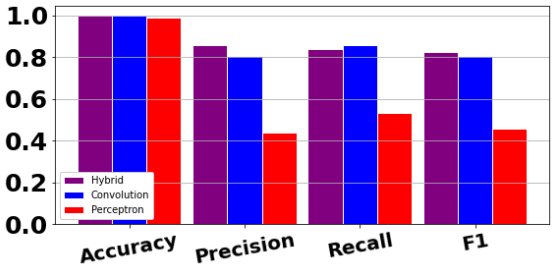
6.3 Comparing Classification Scores within Each Class

Below we present the average accuracy, precision, recall and F1 score for each class, as averaged across 10-folds cross validation, comparing the three architectures. These plots represent a more "micro" representation of classification by showing the response and scores within each specific class.

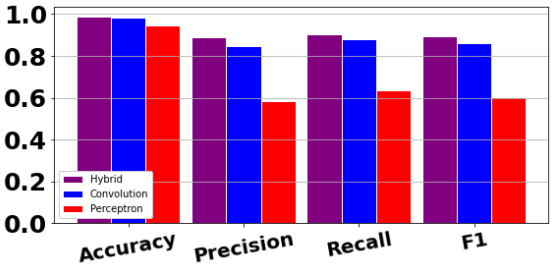
6.3.1 High Woodwind Scores



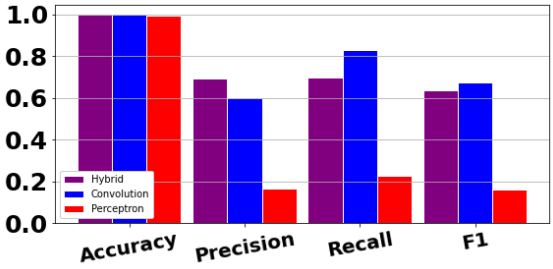
(a) Alto Flute



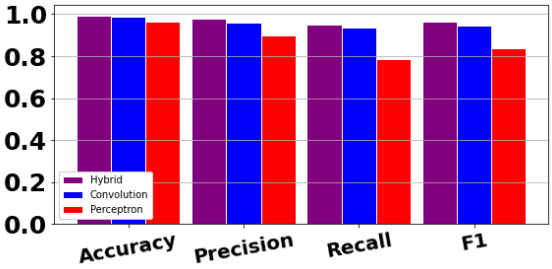
(b) Alto Saxophone



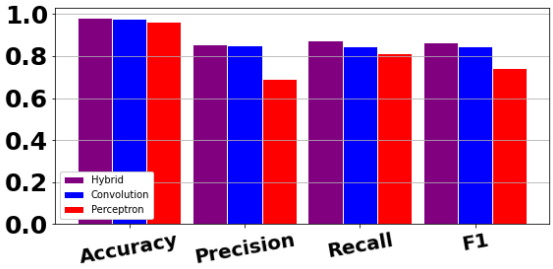
(c) $B\flat$ Clarinet



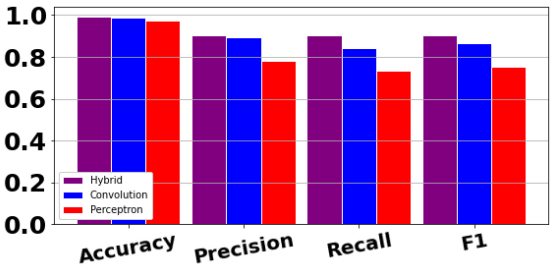
(d) $E\flat$ Clarinet



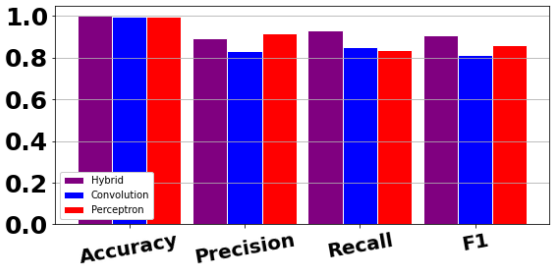
(e) English Horn



(f) Flute



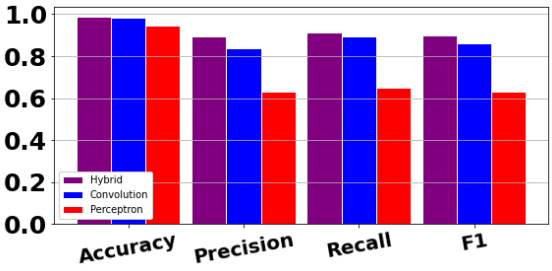
(g) Oboe



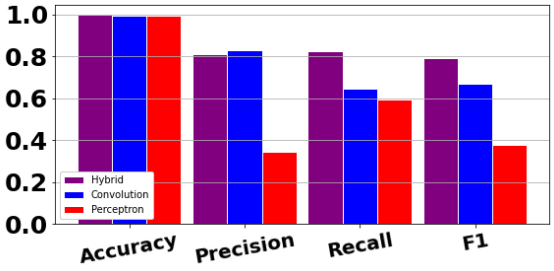
(h) Soprano Saxophone

Figure 35

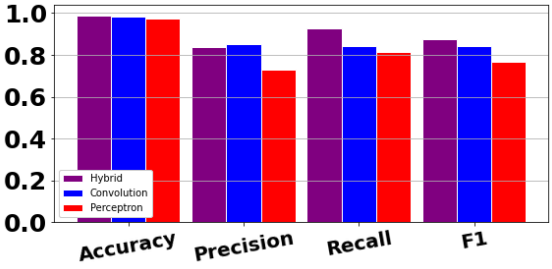
6.3.2 Middle and Low Woodwind Scores



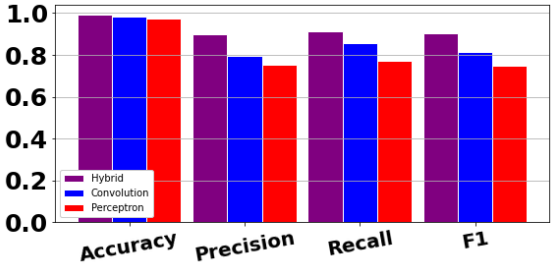
(a) Bass Clarinet



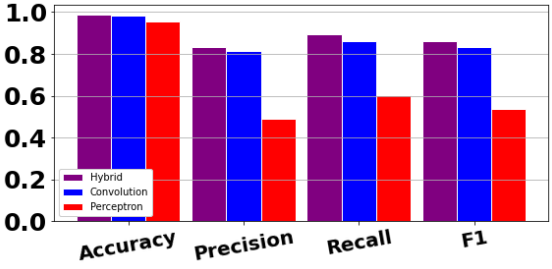
(b) Bass Flute



(c) Bassoon



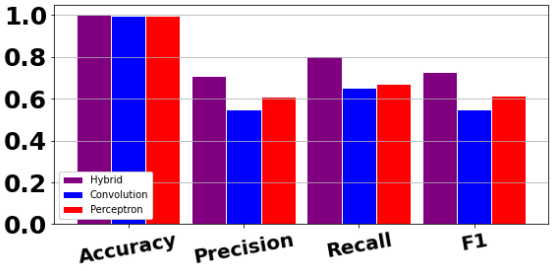
(d) Contra Bassoon



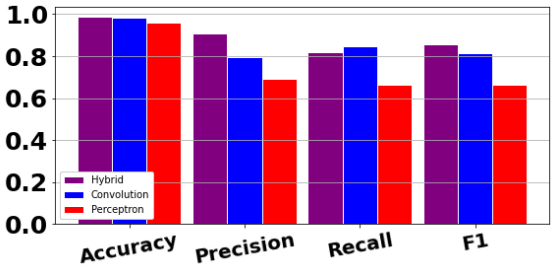
(e) Tenor Saxophone

Figure 36

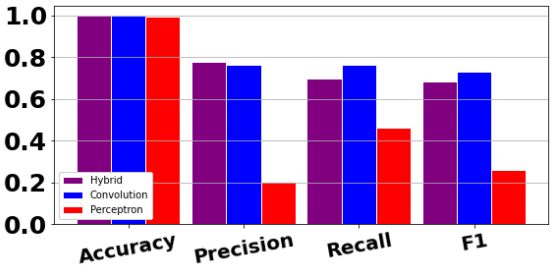
6.3.3 Brass Scores



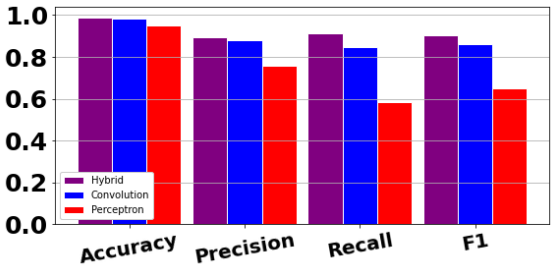
(a) Bass Trombone



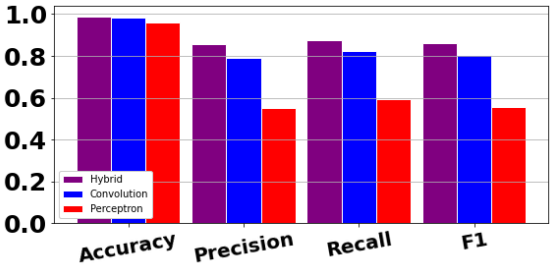
(b) French Horn



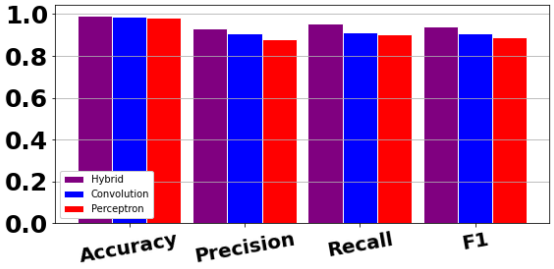
(c) Tenor Trombone



(d) Trombone



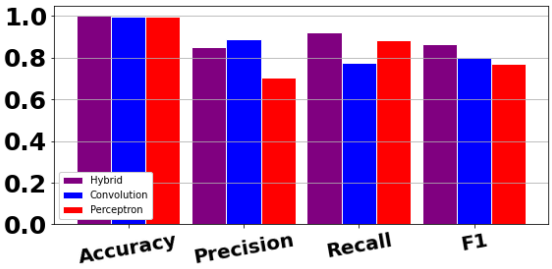
(e) $B\flat$ Trumpet



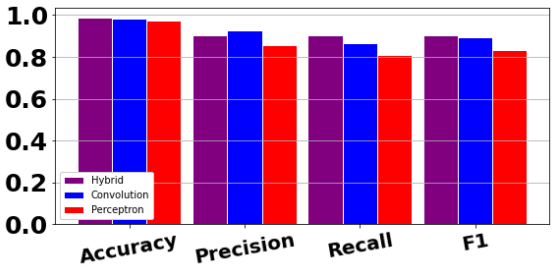
(f) Tuba

Figure 37

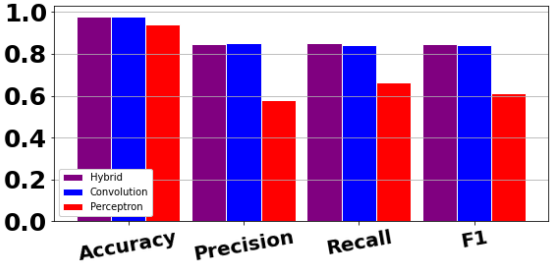
6.3.4 String Scores



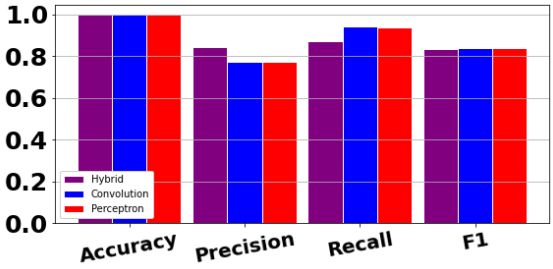
(a) Banjo



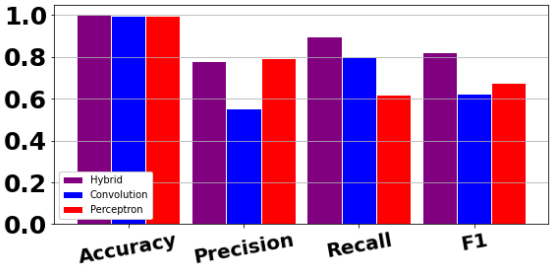
(b) Double Bass



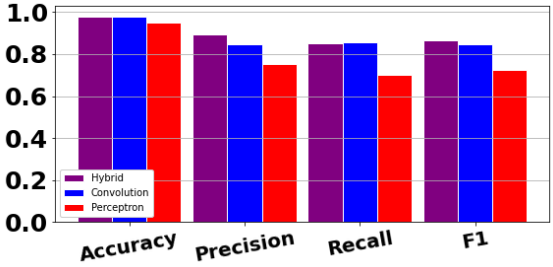
(c) Violoncello



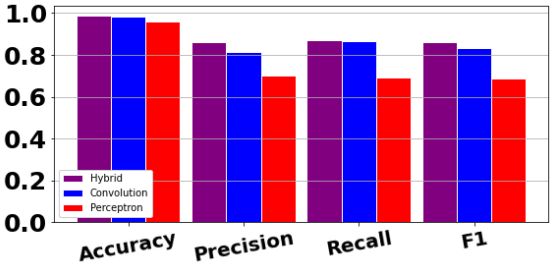
(d) Guitar



(e) Mandolin



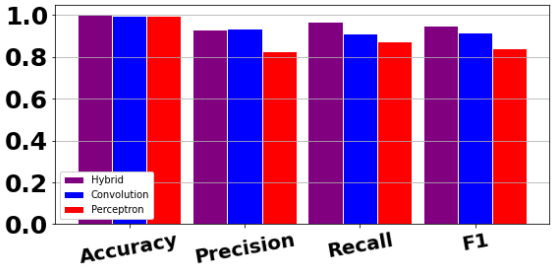
(f) Viola



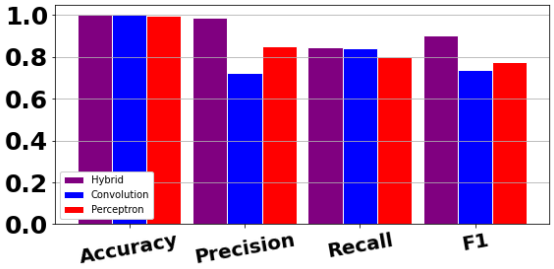
(g) Violin

Figure 38

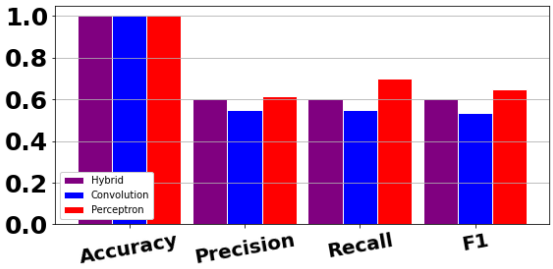
6.3.5 Percussion Scores



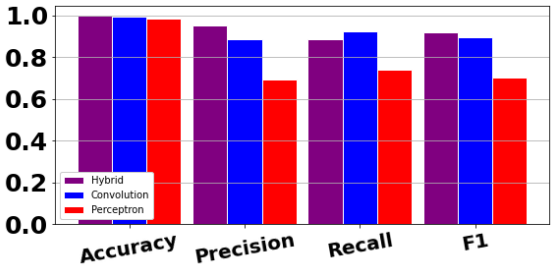
(a) Bells



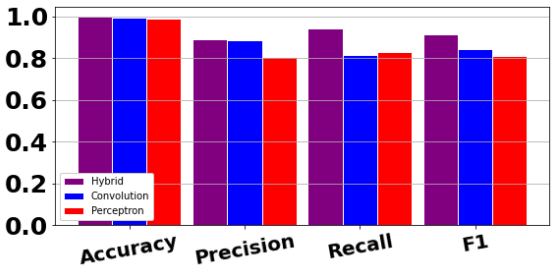
(b) Crotales



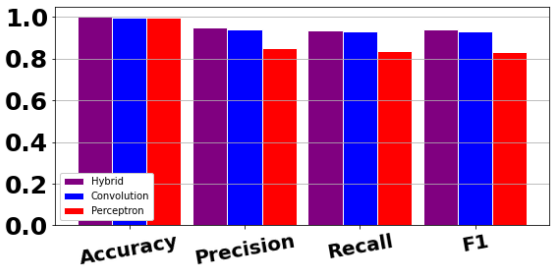
(c) HiHat



(d) Marimba



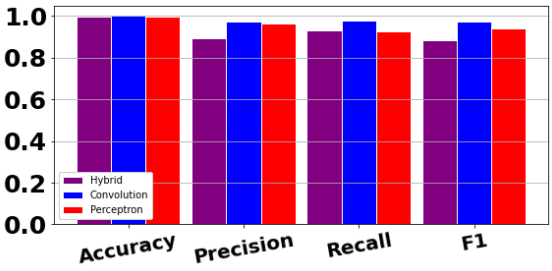
(e) Vibraphone



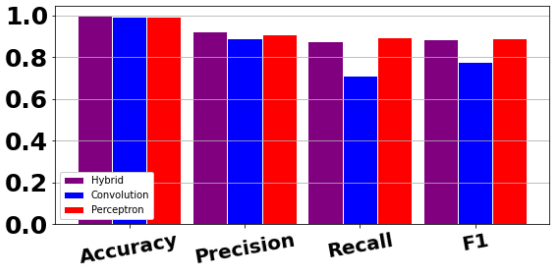
(f) Xylophone

Figure 39

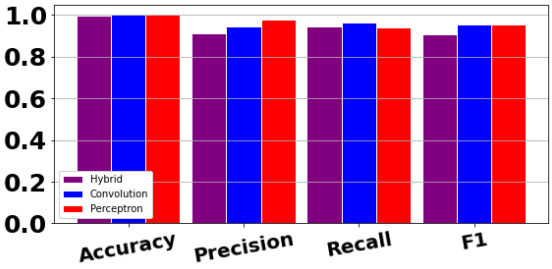
6.3.6 Synthetic Waveform Scores



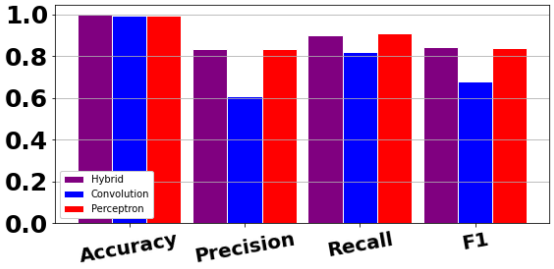
(a) Sawtooth Wave



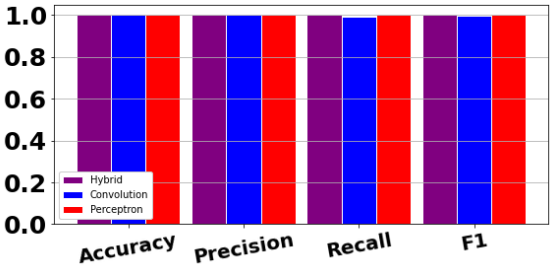
(b) Sine Wave



(c) Square Wave



(d) Triangle Wave



(e) White Noise

Figure 40

6.4 Discussion of Results

Comparing the results within each class, as averaged over 10 folds of cross validation indicates that our model performs very well at the ability to a match a sound waves in digital audio files to their sources. We see this behavior in the plots in section (6.3) as well as the confusion matrices in Fig. (32). For the chosen set of classes, features and architecture, the model can generalize to new information allowing it to distinguish between different types of musical instruments reasonably well for samples that it has not yet interacted with. Additionally, we show that the predictive power of the model improves through our development of the multimodal architecture. By choosing to express the audio file contents in two different modalities of the spectrogram matrix and the feature vector, the hybrid model demonstrates a generally higher classification performance across the same number of samples and splits.

Despite the higher average scores in the hybrid network, see Fig. (31) compared to each unimodal network, Fig. (33), some exceptions do arise at a micro-level (within each class). Consider the Hi-hat instrument in Fig. (39c). Over the cross validations, it was found that the multilayer perceptron architecture produced a better precision, recall, and F1 score than either the convolution or the hybrid network architecture. This can be contrasted with most other classes showing that the hybrid model performed better than either individual mode, with the convolution architecture as a close second.

While the hybrid model can differentiate between many types of instruments, each unimodal variant shows slightly less success over the full data set. In many instances, the MLP-only architecture particularly shows unstable and inconsistent performance. For certain instruments like the $E\flat$ clarinet, tenor saxophone, tenor trombone, and hi-hat, the feature vector does a very poor job of isolating and identifying those classes. However, for instruments such as the soprano saxophone, tuba, and square wave, the feature-vector allows the MLP to perform very reasonably. We see a similar imbalance with the CNN-only architecture where classes such as alto flute, bass flute, bass trombone, and mandolin show poor classification rates. This can be contrasted with English horn, tuba, bells, and double bass, where a much higher classification performance is found. Further exploration is required to provide a full explanation and justification of why this wide range of performance is observed.

We can also observe a loose correlation to the number of samples in each class compared to the classification metric scores. When a model is trained on one class more than another we are more likely to generate predictions for that class too. This can be seen in the scores and occurrences of alto flutes (35a) with 72 samples compared to that of English horns (35e) with 1382 samples. In both confusion matrices of Fig. (32), we see a relatively small entry in $C_{0,0}$ (alto flutes) compared to $C_{14,14}$ (English horn). This indicates that fewer correction predictions were made, and the normalized matrix shows that fewer predictions were made at all. Contrast this with the English horn having a dark red square in both matrices shows many predictions, and many correction predictions. Future work could explore how classification performance would be affected when weighting each class inversely to it's number of samples in the data set. This would allow each class to be weighted uniformly in the training process and may potentially mitigate some of the variability of the performances across classes.

Future work on this project would benefit from a further exploration and development of input features. While the ability to discern classes is present, classification scores are still unstable and we observe a wide range of performances. To combat this, we would produce additional feature that convey characteristic properties of the digital audio files that allow for the generation of more distinctive decision boundaries. In addition to new features, it would be beneficial to further study the distribution of samples in feature space and execute some form of K-best features algorithm. This would indicate which of the chosen predictors consistent or inconsistent with our requirements of intra-class and inter-class variability.

In addition to input features, it may be beneficial to group for the the related classes together. For example, violins and violas have very similar physical constructions, have very similar formant structures, and thus have very similar behaviors in our feature-space. As a result, some humans may also have difficulty differentiating the two, so it may be forgiven that a classification neural network would do the same. We can also consider grouping range variants of parent instruments such as flutes, alto flutes, bass flutes into a single *flutes* category. It is possible that the new distribution of properties in feature-space may also warrant the development of a new set of predictors again.

Finally as a practical demonstration, future work entails the deployment of this classifier to generally unlabeled audio samples, provided that they have the same format. In particular, we would like to explore how classification would behave when testing it on digitally generated chaotic synthesizer wave forms. These signals are representations of chaotic systems that resemble some of the acoustic the periodic properties of digital audio waveforms. By using the automated classifier, we would like to test the versatility of the classifier when subject to samples that may differ drastically in oame ways from the training data set. This demonstration would be a fantastic exploration into the ability for the model to generalize, and the strength of the learned function parameters.

7 Conclusion

The experimental results show that we have successfully constructed a multimodal neural network that can classify sound waves into one of 37 categories that it most resembles. Each category is representative of a particular musical instrument, and collectively the categories cover many stringed, woodwind, brass, mallet percussion, and synthetic instruments. From performing a 10-folds cross validation, we find that given slightly different initial conditions and overlapping subsets of training data, the model produces consistency reasonable classification scores as shown in Fig. (31). This behavior indicates a neural network architecture and features that can generalize to unseen data samples.

Examining the two diagonally dominant confusion matrices in Fig. (32) indicates that the neural network is consistently making correct categorical predictions. The presence of off-diagonal entries in the confusion matrices indicates that misclassifications are a common possibility. Averaged across 37 classes, we see in Fig. (31) that the classification accuracy score averages around 0.98 and precision and recall scores around 0.83 and 0.82 respectively. This means that examining each class in an *one-versus-all* context, the model selects more than 83% of relevant samples, and more than 82% of relevant samples are selected. This indicates that the multimodal neural network that can reasonably differentiate between 37 classes of musical instruments.

We conclude that on average, the hybrid neural network architecture in Fig. (6) exhibits superior classification performance over either of it's unimodal counterpart architectures. In examining confusion matrices from both unimodal models in Fig. (34), we see the presence of diagonal dominance. However, in comparing this to the multimodal network confusion matrix from Fig. (32), we see that the multimodal network has a much more pronounced main diagonal. This indicates that by aggregating to the output of each modal branch with a concatenation layer, we can benefit from the predictive power of both architectures together.

8 Acknowledgments

The completion of this project would not have been possible without the various contributions from the following people and organizations:

- Dr. Kevin Short and Dr. Maurik Holtrop for advising this project.
- Dakota Buell, Tan Dao, Nathan Richard, and Morgan Saidel for consultation on physics, mathematics, and programming topics.
- Madeline Edwards and John Parker for differing me to this project, as well as constant support and consultation on mathematical topics.
- Dr. Kourosh Zarringhalam and Dr. Marek Petrik for additional consultation on machine learning topics.
- University of Iowa, Electronic Music Studio and Philharmonia Symphony Orchestra for the digital sound library used for training data samples.

References

[1] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer New York, 2016.

[2] Geron, Aurelien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly, 2017.

[3] Geron, Aurelien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed., O'Reilly, 2019.

[4] Goodfellow, Ian, et al.*Deep Learning*. MIT Press, 2017.

[5] "Wave Equation: Vibrating Strings and Membranes." *Applied Partial Differential Equations with Fourier Series and Boundary Value Problems*, by Richard Haberman, Pearson, 2019, pp. 130–150.

[6] Von Hornbostel, Erich, and Curt Sachs. "Classification of Musical Instruments." *The Galpin Society Journal*, Translated by Anthony Baines and Klaus Wachsmann, vol. 14, Mar. 1961, pp. 3–29.

[7] Hunter, Joseph L. *Acoutics*. Prentice Hall, 1957.

[8] James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017.

[9] Khan, M. Kashif Saeed, and Wasfi G. Al-Khatib. "Machine-Learning Based Classification of Speech and Music." Multimedia Systems, vol. 12, no. 1, 2006, pp. 55–67., doi:10.1007/s00530-006-0034-0.

[10] Levine, Daniel S. *Introduction to Neural and Cognitive Modeling*. 2nd ed., Routledge, 2000.

[11] Li, Yingming, and Ming Yang. "A Survey of Multi-View Representation Learning." Journal of LateX Class Files, vol. 14, no. 8, Aug. 2015.

[12] Liu, Zhu, et al. "Audio Feature Extraction and Analysis for Scene Segmentation and Classification." Journal of VLSI Signal Processing, vol. 20, 1998, pp. 61–79.

[13] Loy, James , *Neural Network Projects with Python*. Packt Publishing, 2019

[14] John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55

[15] McCulloch, Warren S., and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, 1943, pp. 115–133.

[16] Mierswa, Ingo, and Katharina Morik. "Automatic Feature Extraction for Classifying Audio Data." *Machine Learning*, vol. 58, no. 2-3, 2005, pp. 127–149., doi:10.1007/s10994-005-5824-7.

[17] Mitchell, Tom Michael. Machine Learning. 1st ed., McGraw-Hill, 1997.

[18] Ngiam, Jiquan, et al. "Multimodal Deep Learning." 2011.

[19] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2.

[20] Olson, Harry E. *Music, Physics and Engineering*. 2nd ed., Dover Publications, 1967.

[21] Peatross, Justin, and Michael Ware. *Physics of Light and Optics.* Brigham Young University, Department of Physics, 2015.

[22] Petrik, Marek. "Introduction to Deep Learning." Machine Learning. 20 April. 2020, Durham, New Hampshire.

[23] Philharmonia Symphony Orchestra home page- *https://philharmonia.co.uk/*

[24] Powers, David. (2008). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Mach. Learn. Technol. 2.

[25] Sahidullah, Goutam S. "Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition." 18 Nov. 2011.

[26] Serizel, Roman, et al. "Acoustic Features for Environmental Sound Analysis." Computational Analysis of Sound Scenes and Events, by Tuomas Virtanen, Springer, 2018, pp. 71–101.

[27] Pauli Virtanen, et. al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

[28] Short, K. and Garcia R.A. 2006. "Signal Analysis Using the Complex Spectral Phase Evolution (CSPE) Method." AES: *Audio Engineering Society Convention Paper*.

[29] Fabian Pedregosa, et. al. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)

[30] "Continuum Mechanics." *Classical Mechanics*, by John Robert Taylor, University Science Books, 2005, pp. 681–738.

[31] TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[32] Virtanen, Tuomas, et al. *Computational Analysis of Sound Scenes and Events.* Springer, 2018.

ffl