

Artificial Synesthesia: Exploring the Replication of Synesthetic Responses Using Artificial Intelligence

Nicole Wood

Southern Utah University

Abstract

After observing an individual with sound-to-color synesthesia and recognizing patterns and consistencies, I sought to find if replicating these synesthetic responses artificially was possible. To create a ubiquitous representation of genres, the synesthetic responses of one synesthete (referred to as subject) were documented for 300 songs across 20 genres. These synesthetic responses were classified in three categories: Color, Texture, and Luminosity. Three 30-second samples were created from each song. The Mel Frequency Cepstral Coefficients (MFCC) for each sample were then extracted and used as the model's input.

It was determined using a confusion matrix that the only way for the model to accurately guess classifications was if each category had equivalent representation. The subject selected the best 15 songs in each category to retrain the data, which produced improved results. A success metric of 50% was given to determine if a model could accurately replicate synesthesia. Each response model (Color, Texture, Luminosity) had a higher accuracy than the success metric (50.32%, 58.27%, 65.87% respectively) according to the training/testing dataset. The compiled responses for a given song sample also had an accuracy of 51.2% according to the subject.

With all accuracy scores above 50%, this shows that even with such a small sample, synesthesia can be accurately modeled using artificial intelligence. Using a larger sample size would produce even better results.

Introduction

Synesthesia is the rare phenomenon in which the stimulation of one sense stimulates another sense in the body. The term synesthesia is derived from sym, the Greek root meaning ‘together’, ‘same’, or ‘joined’¹, and aesthesis, meaning ‘sensation’, ‘feeling’, or ‘perception’. Together, synesthesia means ‘joined sensation’². According to Dr. Vivek Ballga³, a synesthete- or person with synesthesia- can experience this phenomenon in a variety of ways. This project will focus on the variation of synesthesia called Chromesthesia.

Chromesthesia occurs when an individual automatically perceives color or visuals to any sound. These non-visual stimuli can be anything from music, speech, phonemes (distinct sounds between letters like p, b, d, t), or everyday sounds.

This phenomenon – and any other form of synesthesia— is caused by the limbic system of the brain. Projections from the limbic system to the neocortex affect our interpretations and models of the world around us². The ‘limbic brain’ is where emotional evaluations take place, which is the driving force for synesthesia. This part of the brain also uses memory structures, which may contribute to the static responses created by synesthesia.

Because these patterns are hard-coded into a synesthete’s brain, an important criterion to be diagnosed with synesthesia is that responses are consistent. For example, if an individual with Chromesthesia hears a song to be green, that song will almost always be green. Although, it is important to note that the synesthete’s emotional state may alter the color of a song slightly.

Synesthesia was reported to be found in 1 in 1,150 adult females and 1 in 7,150 adult males⁴. Most of these individuals do not recognize that they have synesthesia until much later in life, as they assume the rest of the world experiences things the same way they do. However, not all synesthetes with the same type of synesthesia experience the same visual responses. For example, one synesthete with chromesthesia may see a song as red consistently, while another

¹ Eide, “The Root SYM”

² Cytowic, “Synesthesia: Phenomenology And Neuropsychology A Review of Current Knowledge”

³ Ballga, “Synesthesia Types, Causes, Symptoms – Seeing Sounds Explained”

⁴ Rich et al., “A systematic, large-scale study of synaesthesia: implications for the role of early experience in lexical-colour associations”

will always see it as blue. Decades of neuroscientific research have proved that every individual has a unique way of perceiving the world⁵.

After conversing with an individual with synesthesia, I recognized a series of patterns. First, I noticed that certain sounds and genres were always the same color. For example, jazz was usually a luminous yellow or orange, classical was a variant of brown, and similar-sounding pop songs were a sparkly blue. Next, I recognized that over time the songs never changed color (again, a criterion of synesthesia). From what I knew of Artificial Intelligence (AI), this unchanging pattern should be repeatable. If it was, AI could introduce opportunities to allow artificially intelligent models to learn and recognize biometric and limbic patterns in the brain, making it easier to diagnose Synesthesia from a young age.

Methods

Because Synesthesia is different for every individual, the goal of this project was to teach an AI to replicate the synesthetic responses of one young adult woman (who will be referred to as our ‘subject’). The visual responses that she experiences are usually very specific and unique scenes that could be represented with a palette of colors. For simplicity, each visual was reduced to its most prominent color. This meant that although two songs may be labeled blue, the actual visual may be drastically different. To account for this, our subject’s synesthetic responses were broken into three categories: Color, Texture, and Luminosity.

Color will be chosen from 11 options: Red, Orange, Yellow, Green, Blue, Purple, Pink, White, Black, Brown, and Gray.

Texture will be chosen from 7 options: None, Sparkly, Spiky, Woody, Smoky, Smooth, and Soft. Finally,

Luminosity will simply be marked with either Yes or No, indicating the presence or absence of light

According to our subject, “A color that is lighter in shade will have luminosity marked ‘yes’, and the opposite applies for darker colors. Texture is my personal way of interpreting a color overall, as it can be very specific with imagery. While many songs have multiple colors,

⁵ Otis, “Rethinking Thought: Inside the Minds of Creative Scientists and Artists”

for this project I will be focusing on the most prominent one, which is also why texture is important.” As an example, maroon might be coded as Color – Red, Texture – Woody, Luminosity – No.

Each category would be represented by its own model. Each model would be considered ‘successful’ if it predicted the correct response with over 50% accuracy. The overall Synesthetic AI would be considered ‘successful’ if it guessed the combination of each category with over 50% accuracy (according to our subject).

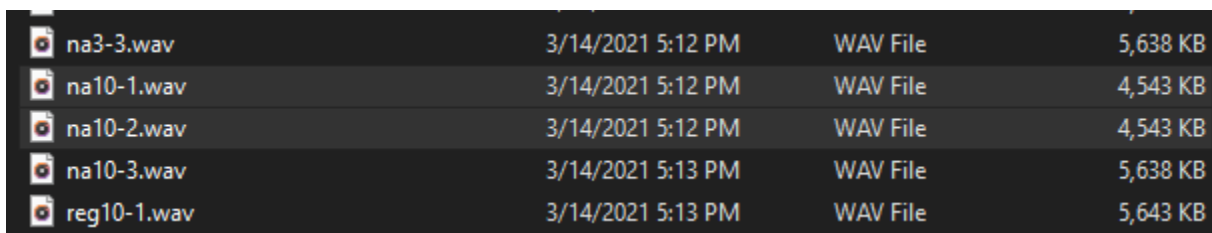
Because all of the data must come from one person, it was decided that – for the scope of this project – 300 songs would be a sufficient representation to determine if synesthesia could be replicated artificially. Although songs in the same genre can have a similar visual, two songs with a common visual are often wildly different genres. Despite recognizing that genres were already capable of being recognized by AI, and colors sometimes favored different genres, I did not want to fit the model to be biased to any genre(s). For example, if I only chose classical songs to represent the Brown category, a brown rap song may generate incorrect results.

To get the most accurate model possible, I wanted to have a wide range of genres represented. I selected an average of 15 songs from 20 genres/categories with distinct sounds: Blues, Classical, Country, EDM, Hip hop, Jazz, K-pop, Lo-Fi, Metal, New Age, Pop (... - 1999), Pop (2000 – 2010), Pop (2011 – 2021), Rap, Reggae, Rock (Alternative), Rock (Classic), Rock (Modern), Rock (Punk), and Soundtrack. For genres with a wide range of sounds (Pop 2011 – 2021), I selected up to 30 songs. For categories that may have some overlapping sounds (Alternative Rock and Punk Rock), I only selected 11-12 songs.

I also did not want to have any model categories (Red, Soft, Yes, etc.) under-represented. However, it can be difficult and time-consuming for a synesthete to seek out a song of a specific color, so I needed to choose a minimum number that was attainable. I decided that with 300 songs and 20 total model categories (11 in Color, 7 in Texture, 2 in Luminosity), each category should have at least 15 songs. After documenting the results for the original 300 songs, the only categories with less than 15 representatives were Pink (in the Color category: 13 songs) and

None (in the Texture category: 12 songs). This meant that our subject only needed to search out 5 songs⁶.

After verifying that each category was sufficiently represented, each song was cut into three 30-second samples. The first would incorporate all or most of a verse; the second would represent the chorus; the third would feature some other notable portion of the song (typically a bridge, key change, guitar solo, or combination of verse and chorus). After the 915 samples were cut, each was verified to be the correct length by checking the file size of each WAV file. Each 30-second song sample was roughly 5,638 KB. Any other file size easily indicated which samples were the incorrect length.



| | | | |
|-------------|-------------------|----------|----------|
| na3-3.wav | 3/14/2021 5:12 PM | WAV File | 5,638 KB |
| na10-1.wav | 3/14/2021 5:12 PM | WAV File | 4,543 KB |
| na10-2.wav | 3/14/2021 5:12 PM | WAV File | 4,543 KB |
| na10-3.wav | 3/14/2021 5:13 PM | WAV File | 5,638 KB |
| reg10-1.wav | 3/14/2021 5:13 PM | WAV File | 5,643 KB |

Figure 1 | Screenshot of song samples and file sizes. Two files are 4,543 KB, indicating that they are not 30 seconds in length.

Samples were moved into 11 folders, each named with the appropriate Color category. Using the names of the folders, the correct mapping, training data, and label index was extracted from each song sample and stored into a data dictionary. After analyzing each sample, this dictionary is outputted to a .json file.

The mapping list will contain the names of each category; these are extracted from the folder names. For the Color model, this would be Black, Blue, Brown, etc.

The sample's Mel Frequency Cepstral Coefficients (MFCC) were used as the model's training inputs. An MFCC captures the textural and timbral aspects of sound (the feeling and quality of music you can decipher from different sounds or instruments) in a format that can be interpreted by computers. Because of this, MFCCs are useful for AI models designed to imitate the human auditory system. Possible applications include speech recognition, music genre classification, and instrument classification.

⁶ Resulting 305 samples, as well as their results, Appendix A.

The label (as an integer) represents the training output. This is the mapping index of the correct label assigned to that sample. For example, since folders are listed in alphabetical order, each sample contained in the Black label would have the (color) mapping index of 0. The next label—Blue—would have a mapping index of 1. Brown would be 2, and so on.

This dictionary is loaded as training data into the model. This model was built using the `keras.Sequential` library, and composed of an input layer, three hidden layers, and an output layer. Regularization and layer dropout were used on each hidden layer to assist with overfitting⁷. The output layer used a softmax activation function for multi-classification. After training, the model's accuracy was ~18%. Given the 11 possible categories, randomly guessing should generate an accuracy of 1/11 or 9.09%. At first glance, this indicated that the model was doing well.

To get a better understanding of what the model was predicting, a confusion matrix was generated using the `matplotlib` and `sklearn` libraries. A confusion matrix allows us to visualize what is being guessed compared to the actual answer. Ideally, you will see the matrix diagonal stand out, indicating that it usually guesses correctly. However, in my first confusion matrix, we can see that the model was guessing Blue almost every time. After looking into the data, I found that 54 songs were labeled Blue. Of the 305 total songs, these 54 songs make up about 17.7% of the data. Therefore the accuracy seemed to be better than random guessing.

⁷ Guides for creating an Audio Classifying AI model courtesy of Valerio Velardo – The Sound of AI on YouTube.

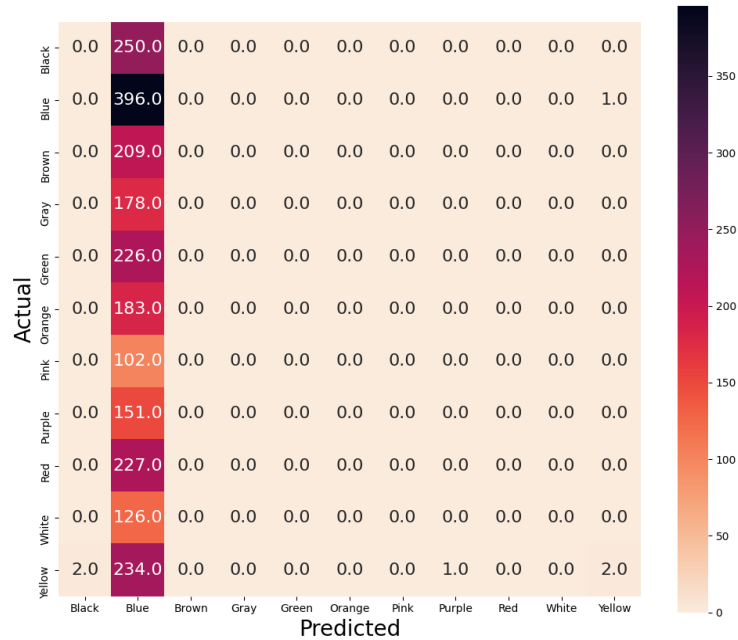


Figure 2| Confusion Matrix indicating the Blue category was chosen for all but 6 samples.

Given that the Blue category had over 20 more songs than any other, I ran the model a few more times without the Blue folder. These runs proved that the model would only guess whichever category had the most songs representing it. In this case, the model would choose either Black or Red, which both had 32 songs.

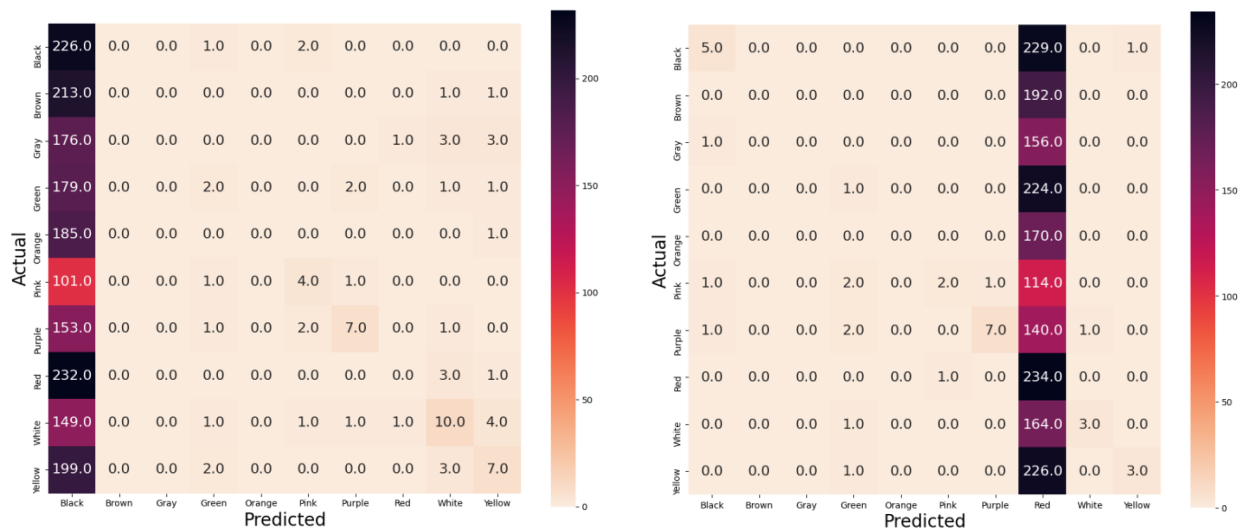


Figure 3| Two separate Confusion Matrices indicating the model is favoring one category above the others.

I determined that if all samples were equally represented, there would not be an obvious category to choose every time. I asked my subject to select the 15 songs that best represented each color from my small dataset.

With 15 of each color, the model was finally starting to show some promise. Although it still chose one color (Brown) to guess most of the time, we were finally starting to see the matrix diagonal light up with correct answers. After a few attempts of playing with the training rate, dropout probability, and regularization, I finally found that increasing the number of epochs (iterations over the training data) improved the results.

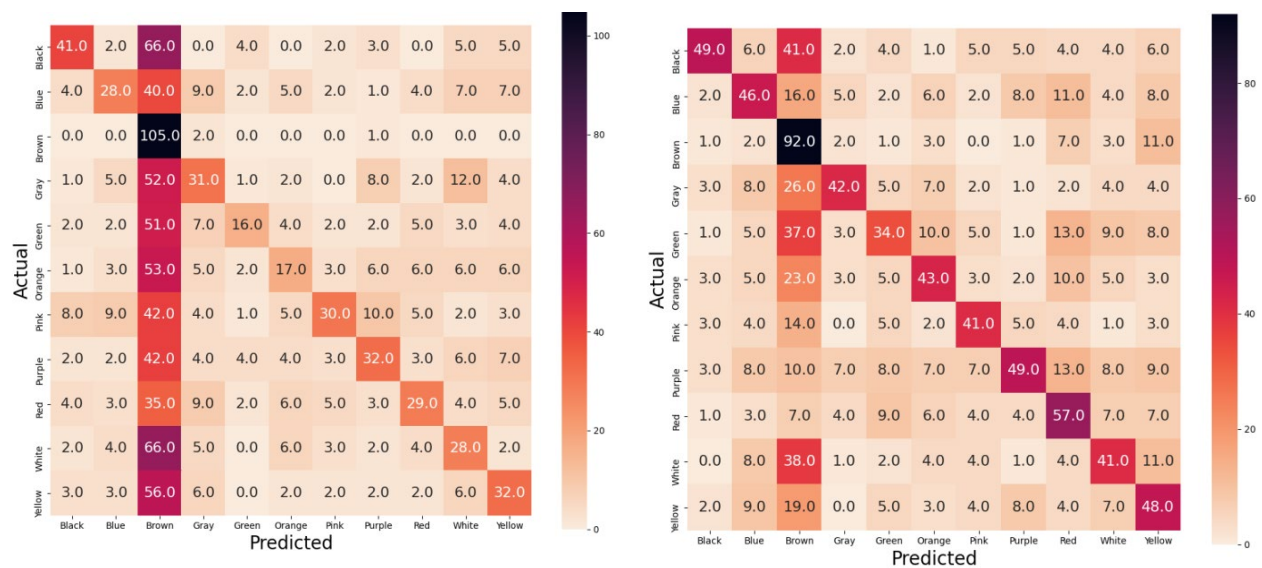


Figure 4| Confusion Matrices after standardizing the number of songs in each category. Increased epochs emphasize the matrix diagonal, indicating more accurate results.

I found that the results did not get much better past 300 epochs, so I kept this number for the Color and Texture models.

Discussion

Using only 15 songs per category at 300 epochs, the Color model outputted an accuracy of 50.32%.

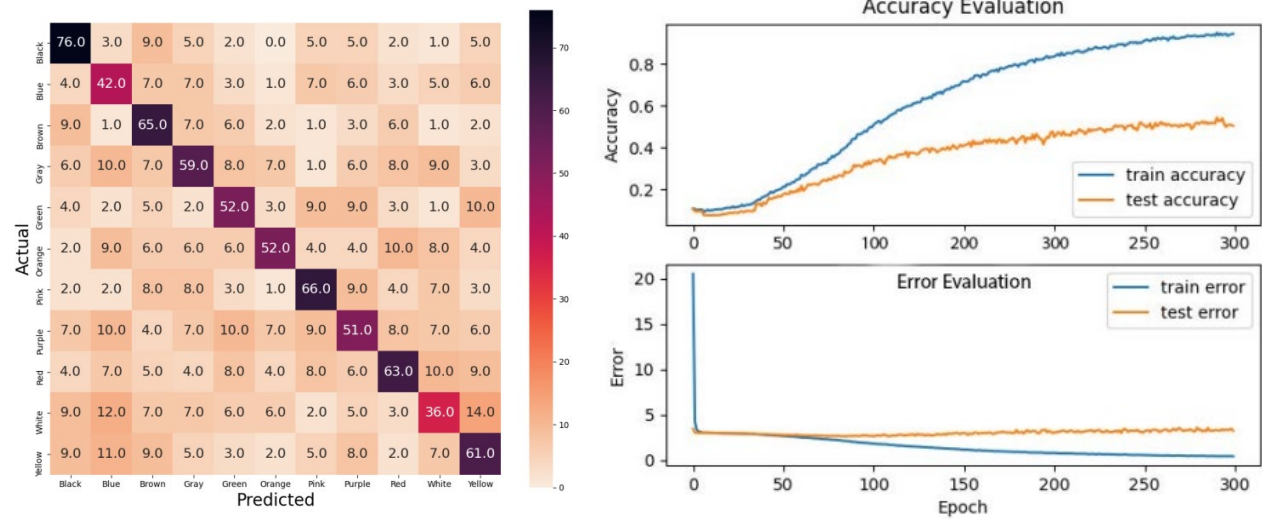


Figure 5| Resulting Confusion Matrix and Accuracy Evaluation for the Color Model

After finding success with the first model, the subject was asked to find the 15 most accurate representations for each Texture category. Using this data, the Texture model outputted an accuracy of 58.27%.

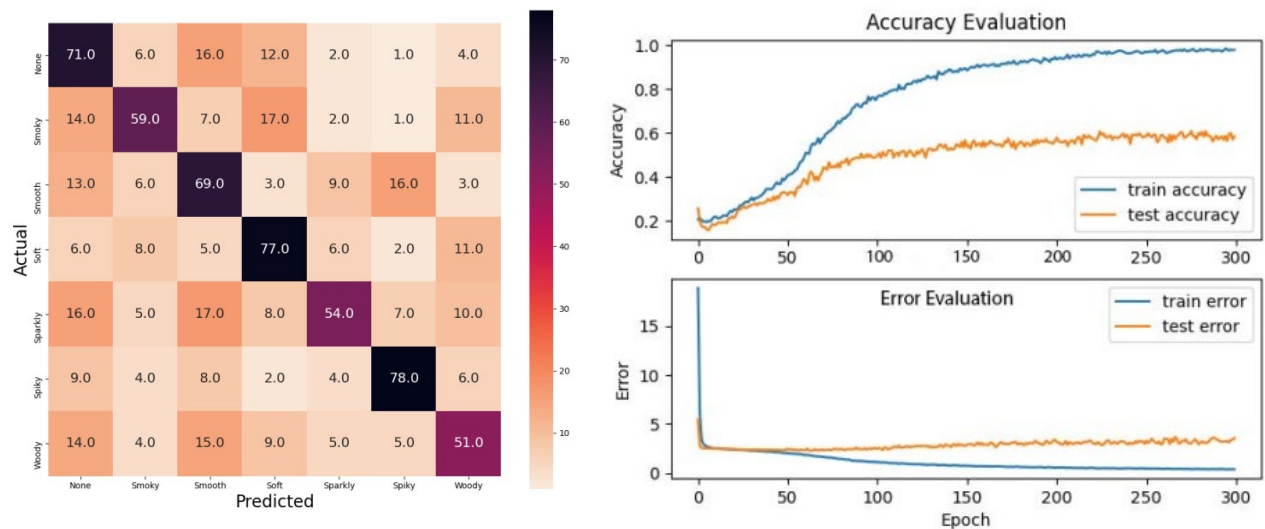


Figure 6| Resulting Confusion Matrix and Accuracy Evaluation for the Texture Model

Finally, with fewer categories to choose from, fewer epochs were required to get acceptable results from the Luminosity Model. With 75 epochs, the Luminosity outputted an accuracy of 65.87% accuracy.

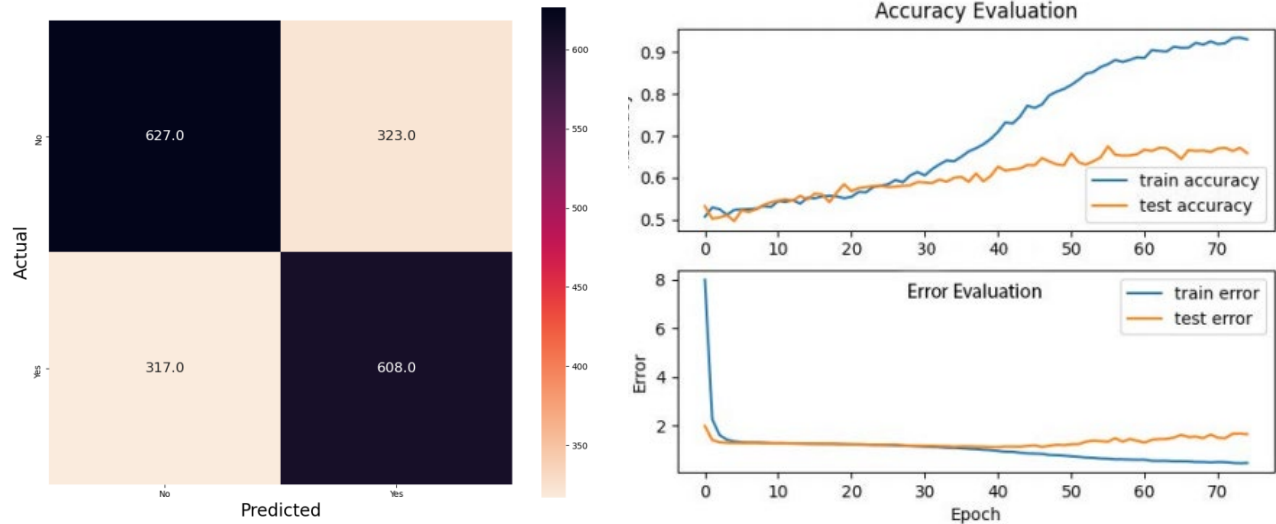


Figure 7| Resulting Confusion Matrix and Accuracy Evaluation for the Luminosity Model

After each model successfully met the minimum accuracy score, a small application was written to compile each model’s prediction into one response for a single testing sample. This response was outputted to the command line for the user to read as “Song is predicted to be [Texture] [Color] [with/without] luminosity”.

```

-----
2021-04-07 23:17:02.911231: I tensorflow/compiler/mlir
Song is predicted to be Soft Brown with luminosity

```

Figure 8| Example of the compiled response to the user

Using this application, predictions were gathered for 25 song samples that were not a part of the original training/testing set. These predictions were given to our subject, who rated the prediction on a scale from 0 to 10 (0 being completely inaccurate, 10 being completely accurate). The average score for these 25 test samples was 5.12 (51.2% accuracy according to our subject).

Unfortunately, given the scope and time limit for this project, we were forced to use an extremely small dataset. I would have liked to add songs to our dataset to bring each category up to at least 54 samples each, as this would maximize the use of our data. A dataset of 305 total songs was already too small, but because the AI did not work when categories were not equally represented, I had to throw out 54% of my data just for the Color model.

The subject gave me a time estimation of two hours to find 20 songs of a specific color⁸. She also mentioned that it would be likely that each song she found would be of the same genre since it is easy to find songs that sound similar to each other. Because I made it a priority of finding a range of genres, I would either have to sacrifice the even distribution of genres, the time spent to find new songs, or the number of samples for each category.

Conclusion

Even with such a small dataset, each model successfully met our 50% minimum accuracy score, and the overall Synesthetic AI received an accuracy score over 50% according to our subject. Even with such little data, the model's accuracy scores were impressive. This proves that with enough data, Synesthesia *can* be replicated artificially. If this technology were to be explored and expanded, it might be possible for neuroscientists to use AI to research and better understand this neurological phenomenon without the limitations of using a human subject.

⁸ Subject synesthete in discussion with the author, April 2021

References

- Cytowic, Richard E. (1995). "Synesthesia: Phenomenology And Neuropsychology A Review of Current Knowledge". PSYCHE 2(10). <http://psyche.cs.monash.edu.au/v2/psyche-2-10-cytowic.html>
- Otis, Laura. (2015). "Rethinking Thought: Inside the Minds of Creative Scientists and Artists". Oxford University Press. <https://doi.org/10.14507/er.v23.2082>
- Eide, Denise, Logic of English. (2012, November 29). "The Root SYM". [Logicofenglish.Com](http://Logicofenglish.com).
- Rich, A.N et al. (2005) "A systematic, large-scale study of synaesthesia: implications for the role of early experience in lexical-colour associations". Cognition 98(1):53–84. doi:10.1016/j.cognition.2004.11.00

Appendices:

Appendix A:

Data and Responses:

<https://docs.google.com/spreadsheets/d/1HJfHXcbVR-pNbT7ZG6pJSjnzLiJLgVJDMIFB4BtXrLc/edit?usp=sharing>