Identifying Music Genre from Electroencephalogram Readings, and Optimising Neural Network Performance Using Prediction Thresholds and Genetic Algorithms

C.M. McCluskey

Research School of Computer Science, Australian National University
U4224263@anu.edu.au

Abstract: Electroencephalography (EEG) has been used to monitor the effects of music on brainwaves and emotions, which has significant implications in neuroscience in using music as therapy. However, EEG data can be difficult to interpret given the magnitude of data collected and the lack of homogeneity between individuals. This paper looks at how neural networks can be used to analyse and interpret EEG data, by building a neural network to classify music genre from EEG data. It also looks at how neural networks can be optimised for different purposes by changing the prediction threshold and by using a genetic algorithm for feature selection to reduce the magnitude of data being utilised by the model. It is found that changing the prediction threshold for a neural network can alter the precision and recall of the network without compromising network performance. It is also found that a decently performing binary classifier neural network (69.92% accuracy) can be built to classify music genre, which is further improved by using a genetic algorithm to select relevant features (73.77% accuracy), however further work is needed to try more complex modelling using more data.

Keywords: EEG; neural network; prediction threshold; genetic algorithm; feature selection; music therapy.

1 Introduction

Music has, throughout history, been of vast cultural significance, and has been used as a form of therapy by many cultures, spanning from ancient to modern times. Historically, the role of music as therapy has been shaped by religious, cultural and social effects, but going into the twentieth century it has become a focus of neuroscience to see the effects that music can have physically on the brain (Thaut, 2015; Rahman et. al., 2020).

A well-known way to measure what is going on in the brain is using electroencephalography (EEG) to measure electrical signals. This is done by placing electrodes on the scalp and measuring signals coming from different areas of the brain for a specific time frame (Fachner and Stegemann, 2013). Different signals received by the EEG represent different emotional states within the brain (Fachner and Stegemann, 2013; Rahman et. al., 2020). This is important because music therapy is an often used form of therapy for treating depression, anxiety, attention disorders, and even epilepsy and seizures, among other medical issues. However, the type of music used for therapy may be important. Some genres may be more effective than others, or more suited to specific purposes, or music may have to be tailored to an individual treatment plan. In fact, EEGs can record similar effects in different individuals, despite the genre of music being listened to, due to individual preferences (Wilkins et. al., 2014). Furthermore, information gained from EEGs has been found to be poorly homogenous between individuals, making modelling using this data difficult (Fdez et. al., 2021). Therefore, understanding how music affects individuals, and which kinds of music are likely to cause improvement in specific disorders is vital to continue implementing effective treatments (Fachner and Stegemann, 2013; Rahman et. al., 2020). Being able to analyse this information to provide meaningful answers as to what kinds of music are the most effective therapies is therefore very important.

Another difficulty with using EEG data is that it has high-dimensionality. EEG data is often collected from multiple channels which come from various electrodes attached to the subject's head. Each channel also records multiple features of information (Fachner and Stegemann, 2013). This high dimensionality means that models using EEG data may often include features that are redundant for the classification problem being explored, which can result in the model overfitting (Nakisa et. al., 2018, Sabeti et. al. 2007). However, many feature selection methods offer limited

improvements to the model because they assume that features are independent from one another. This means that features may be disregarded as irrelevant, even though they may be relevant when combined with other features as well as the opposite problem of keeping features that are redundant (Nakisa et. al. 2018). One proposed way to overcome this issue is through the use of genetic algorithms combined with neural networks which can analyse the whole population of features together by iterating through subsets of the features to improve the selection based on selected optimization criteria (Nakisa et. al. 2018, Sabeti et. al. 2007).

This paper investigates the use of neural networks to process EEG information, using a subset of data sampled by Rahman et. al. (2020), to try to predict what genre of music was being listened to based on statistical features derived from EEG data. The genres of music included in the data were classical, instrumental and pop. The neural network was trained to predict these multi-class genres, as well as to predict a binary classification of classical/instrumental and pop. This paper additionally investigates whether a neural network can be optimized by altering the prediction threshold of the network, following Milne et. al. (1995). This is important because it is typically assumed that a prediction threshold of 0.5 per cent is optimum but Milne et. al. (1995) proposed that this threshold should not be 'sacrosanct' and found that changing it affects the number of false positive relative to false negative results returned by a neural network. Finally, the paper explores whether genetic algorithms can used to refine the feature selection from the EEGs to reduce the number of redundant features selected and thus optimize the neural network.

2 Method

The dataset used in this analysis is a subset of a larger dataset that was collected and used by Rahman et.al. (2020). The original study collected electroencephalogram (EEG) data using an Emotive EPOC headset. The headset was used to record EEG data from 24 participants while they listened to several musical pieces that were either classical, instrumental, or pop genres. The headset recorded 14 channels of data from various electrodes on the participant's head while the participant listened to the music. The time series data from each channel was then extracted to 26 linear and non-linear statistical features for each piece of music that a participant listened to (Rahman et. al., 2020). The dataset used for this analysis contained only information from the F7 channel and 25 associated statistical features as this was the data that was available at the time of the analysis, therefore only information from one part of the brain was represented. The 25 linear and non-linear statistical features for this channel are shown in Table 1. The dataset contained 576 rows of data with each row representing a record of a subject listening to a particular piece of music, the 25 features described in Table 1, a participant identification number, and music genre. The numbers were all floats, but were converted to integers to ensure they were all the same.

Table 1. Linear and non-linear statistical features in the dataset, EEG channel F7

Linear	Non-Linear
Mean, Maximum, Minimum, Standard Deviation, Interquartile Range, Sum, Variance, Skewness, Kurtosis, Root Mean Square, Average of the power of signals, Peaks in Periodic Signals, Integrated Signals, Simple Square Integral, Means of the absolute values of the first and second differences, Log Detector, Average Amplitude Change, Difference Absolute Standard Deviation Value	Detrended Fluctuation Analysis, Approximate Entropy, Fuzzy Entropy, Shannon's Entropy, Permutation Entropy, Hjorth Parameters, Hurst Exponent

Source: Rahman et.al. (2020), p. 4 (of downloaded document).

Inspection of the data using basic statistical summaries revealed big differences between the values of the 25 features. The mean of each feature is plotted in Figure 1 as an example of the difference. Large differences we also seen in maximum and minimum values and standard deviation. Based on this information the Fuzzy Entropy and the Simple Square Integral features were dropped from the dataset prior to modelling as these would significantly skew modelling results given the magnitude of the difference between these features and the others. The participant identifier number was also dropped as it was not desirable for the model to learn how each participant reacted to the music, but rather how the general population would react. Furthermore, using EEG signals to represent emotions has been shown to have poor homogeneity between individuals (Fdez et. al., 2021). Therefore, keeping the individuals in is likely to skew the results from the model. However, it should be noted that individuals have been shown to react very differently to different types of music depending on their musical taste (Wilkins et. al., 2014), so the individuals could be relevant for some modelling purposes. Some significant differences still remained in the remaining features, and therefore these were normalised, standardising using the mean and standard deviation.

The data was seeded to ensure replicable results, then shuffled and split into training and test datasets (using an 80/20% split), and input and output features. Then a one hidden layer neural network was constructed, loosely following

Rahman et. al. (2020). However, there were some differences, including that backpropagation was used instead of Levenberg-Marquardt as this was considered easier for the analysis. The dataset was cut in two different ways to train the neural network. The multi-class classifier, following Rahman et. al. (2020), was designed to predict whether the music being listened to was classical, instrumental or pop based on the input features, the other was a binary classifier, where the classification was between classical/instrumental or pop. This classification was chosen because, subjectively speaking, instrumental and classical music have more in common with one another than with pop. Adam was chosen as an optimizer because it performed better than stochastic gradient descent (SGD).

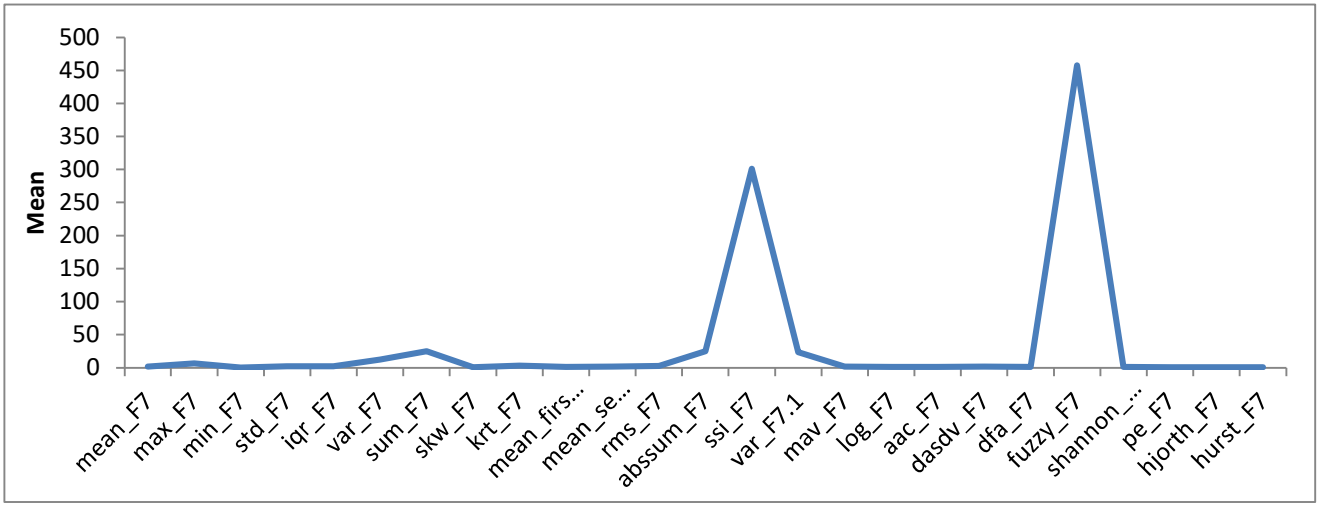

Fig. 1. Mean of the 25 features in the dataset

Additionally, for the binary classifier, the prediction threshold, which is generally assumed to be 0.5, was able to be altered following Milne et. al. (1995). Milne et. al. (1995) found that changing the prediction threshold could alter the number of false negatives and false positives, without necessarily having a significant impact on the performance of the model, thus this effect was tested on the binary model.

For both models, several iterations were run to test the effect of changing the number of epochs, the number of hidden neurons and the learning rate, as well as the prediction threshold for the binary classifier, to build the best performing neural network. Iterations were made between 5 and 20 hidden neurons and 100 to 1000 epochs, with learning rates between 1 and 0.001. The numbers were changed at a fairly course level (by 5s for the hidden neurons and by 100s for the epochs) and then increasingly granularly as the model performance improved.

After running these models, a genetic algorithm was used on the best model (which was the binary classifier with the hyperparameters specified in the results section below) to see if that model could be improved further by selecting the best features and only including those in the model. All features were included in the initial data, excluding the participant identifier number for the reasons described above. The genetic algorithm used mutation as the search algorithm, as this was decided to be simpler than using a crossover function. Accuracy was used as the fitness function to determine the likelihood of selecting each selection of features for the next generation in the algorithm, also chosen for simplicity's sake. The population size, mutation rate and the number of generations were the hyperparameters that were changed at a fairly course level (in 5s and 10s) to see changes in performance. A mutation rate of 0.05 was selected initially (following Sabeti et. al. 2007), with variations from 0.002 to 0.5 to test the changes. Number of generations and population sizes of between 5 and 50 were tested.

3 Results and Discussion

The multi-class classifier performed less well than the binary classifier. This does not seem surprising, since the binary classifier is inherently simpler, and therefore likely to perform better in a simple model. Furthermore, the preferences of an individual to a certain type of music have been shown to have a significant effect on their reactions to the music, so the performance of the binary classifier may also represent a relationship between preferences of individuals to pop versus instrumental and classical music (Wilkins et. al. 2014).

The multi-class classifier was found to perform most optimally with 10 hidden neurons, a learning rate of 0.01 and 650 epochs, noting that not every possible combination was tested. With these parameters a maximum testing accuracy of 42.62% was achieved. This result was significantly lower that than that found by Rahman et. al. (2020), however the models diverged significantly in that Rahman et. al. (2020) used the full dataset rather than only one channel, used an average of 20 runs, which was not done here, used 30 hidden neurons based on the results of a previous study, and used a different network training function (Levenberg-Marquardt) instead of backpropagation.

The binary classifier was found to perform most optimally with 15 hidden neurons, a learning rate of 0.01 and 700 epochs, again noting that not every possible combination was tested. With these parameters an accuracy of 69.92%, precision of 72.34%, and recall of 86.08% was achieved with the threshold set at the standard of 0.5.

A higher accuracy of 71.54% was achieved with a prediction threshold of 0.8, and precision of 85.48% however recall declined to 67.09%. Further details of the effects of changing the threshold on the performance of the network is in Tables 2 and 3. In training, the 0.5 threshold delivered the most correct results, with a fairly even distribution between false positives and false negatives. However, in testing, the total number of correct results was the same at thresholds 0.5, 0.6 and 0.7, and increased slightly for 0.8.

Table 2. Performance of binary neural network on the training dataset, adjusting the prediction threshold

Threshold	Correct (TP+TN)	FP	FN	Incorrect (FP+FN)
0.3	396	54	3	57
0.4	406	38	9	47
0.5	416	22	15	37
0.6	409	13	31	44
0.7	401	8	44	52
0.8	380	3	70	73

Table 3. Performance of binary neural network on the test dataset, adjusting the prediction threshold

Threshold	Correct (TP+TN)	FP	FN	Incorrect (FP+FN)	Accuracy	Precision	Recall
0.3	81	33	9	42	65.85	67.96	88.61
0.4	80	32	11	43	65.04	68.00	86.08
0.5	86	26	11	37	69.92	72.34	86.08
0.6	86	22	15	37	69.92	74.42	81.01
0.7	86	16	21	37	69.92	78.38	73.42
0.8	88	9	26	35	71.54	85.48	67.09

There were fewer false positives and more false negatives as the threshold increased in both the training and the testing data, which is the same as the result observed by Milne et. al. (1995). The implications for this, as noted by Milne et. al. (1995), are that the optimum performance of the neural network will depend on whether it is more important to have fewer false positives or fewer false negatives, or an even distribution between them. This would depend on the what the model was trying to predict. For example, with medical diagnosis it could be more important to have fewer false negatives (so as not to miss a serious diagnosis), but with approval of medicines to market it could be more important to have fewer false positives to ensure only effective medicines make it to market (Ravenzwaaij and Ioannidis, 2019). In this case, since we are trying to predict what type of music a person is hearing based on the effect it has on their EEG, neither false positives or false negatives are preferred, and therefore a threshold between 0.5 and 0.6 provides the optimum results.

The binary classifier above was also supplemented with a genetic algorithm to select features from the dataset, with variations in population size, generation number and mutation rate varied between runs, which produced consistently better results than the simple binary classifier. Table 4 shows the effect of varying these parameters on the accuracy of the model, and which features were selected each time. Best results were achieved with a population size of 10 and 50 generations using a mutation rate of 0.5 which achieved an accuracy of 74.59%. However, this took a significant amount of time and was not significantly better than the accuracy of 73.77% which was achieved in multiple different runs. Many of these runs additionally chose the same features. Therefore, the model which provided the shortest run time while achieving this slightly lower accuracy was preferred. This genetic algorithm had a population size of 10, 10 generations, and a mutation rate of 0.05.

The features selected by this algorithm included minimum, variance, skewness, kurtosis, mean of absolute difference of the first difference, root mean square, simple square integral, average amplitude change, and difference in absolute standard deviation. However, it should be noted that these would not necessarily be the optimum features to model for every EEG problem and the results do not necessarily translate across datasets.

The population size and number of generations seem surprisingly low. However, the dataset did not contain a significant number of features, and, further, this is in line with findings from Chen et.al. (2012) who found that, depending on the data, a higher population size does not always improve results and can actually make results worse. It should also be noted that there was not a significant variation in accuracy with movement in any individual hyperparameter, with a range of only 72.13% to 74.59%. Furthermore, the fact that similar accuracies and features occur despite changes in parameters could indicate that the genetic algorithm quickly moves to find a local maximum and therefore changes in any one parameter do not have a significant effect.

Table 4. Performance of binary neural network on the test dataset, adjusting genetic algorithm feature selection parameters

Population size	Generation number	Mutation Rate	Accuracy	Features
30	15	0.05	72.95	0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 0 0 1 0 0 0
10	15	0.05	73.77	0 0 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 0
50	15	0.05	73.77	1 1 0 1 1 1 1 0 1 1 0 1 0 0 1 1 0 1 0 0 1 0 0 0 0
5	15	0.05	72.13	1 0 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 0 0 0 0 1 1
10	15	0.002	72.95	0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 0 0 1 0 0 0
10	15	0.1	73.77	1 0 0 0 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 0 0 1 1 1 0
10	15	0.5	72.95	0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 0 0 1 0 0 0
10	5	0.05	72.95	0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 0 0 0 1 0 0 0
10	10	0.05	73.77	0 0 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 0
10	30	0.05	73.77	0 0 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 0
10	50	0.05	73.77	0 0 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 0
50	50	0.5	74.59	1 1 1 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0
10	50	0.05	73.77	0 0 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 0
10	50	0.5	74.59	0 1 0 1 1 0 1 1 1 0 0 1 1 0 1 0 1 0 0 0 1 1 0 0 0

Note: Highlighting shows the runs with the highest accuracy and lowest values in the hyperparameters.

Some limitations of this analysis include that the method for optimizing the parameters for the model was fairly course, with a person required to change the inputs by a certain amount rather than automatically iterated through. This could have been improved by writing code that would iterate through possible combinations of parameters, and may have resulted in a better performing neural network. A related limitation is that hyperparameters were testing independently of one another, whereas changing the various hyperparameters together could have produced different results. For example, a model that changed the hyperparameters of both the neural network and the genetic algorithm at the same time, rather than independently, may have resulted in a more optimal model.

A further limitation is that all values represent a single run, rather than an average. This was because the data was seeded to ensure replicability. However, with different cuts of the training and testing data, different results would be observed to those presented here. Running multiple iterations with each of the parameters with different cuts of the data and taking averages might have given a better idea of the optimum parameters for the neural network.

The model presented here is also fairly simple. A more complex model might perform differently, and the results from this model are not necessarily directly comparable to modelling EEG data with other models. For example, modelling that tested further optimizers, other than Adam and SGD, might produce different results, and neural networks that had more hidden layers. Accuracy may also not have been the most appropriate fitness function for the genetic algorithm, and looking at using other fitness functions may have improved results.

The dataset used in this analysis was also fairly small, based on the readings from only one channel in one set of data collection, and based on only 24 participants, limiting the scope of the model to learn and classify. The dataset also only contained music from three genres, out of dozens of potential genres that could be modelled and analysed. The results presented here also do not necessarily translate between datasets.

Lastly, although the model performed fairly well as a binary classifier, it should be noted that individual preferences to music and individual measurements from EEG data can vary widely (Wilkins et. al. 2014; Fdez et. al., 2021). The neural network presented here does not include modeling of the individuals, which is something that may be important for future research.

4 Conclusions and future work

This paper presents a neural network trained to classify music into genres based on EEG readings taken while individuals were listening to different genres of music. A reasonable accuracy was achieved using a binary classifier; however, a multi-class classifier did not produce very satisfactory results. The binary classifier was also further improved by using a genetic algorithm for feature selection. It was also found that altering the prediction threshold could influence that accuracy, as well as the number of false positives and false negatives that the model produces, allowing the model to be optimized to produce either fewer false positives, or fewer false negatives, while maintaining the same number of correct classifications. However, altering the hyperparameters of the genetic algorithm was found to only have minor effects on the accuracy of the binary classifier.

There are several implications from this work. Firstly, the ability to classify music based on EEG signals will enable further analysis in this area, which will be able to inform decisions about how and what kind of music can be used in music therapy. Secondly, the work on genetic algorithms paves the way for further study in this area to determine which EEG signals are the most significant ones to study. Thirdly, the way the prediction threshold can be changed to affect the recall and precision of the model has implications for building neural networks where the number of false positives or false negatives is important, such as in medical diagnosis and drug approvals.

The analysis presented in this paper has some limitations such as the simplicity of the model, the limited number of runs that the model was used for, limitations due to the small size of the dataset used here, and the inability of the neural network to take account of individual influence on the EEG readings. Further work in this area should focus on expanding the modelling to use more complex models and techniques, particularly multi-class classifiers, doing more thorough testing of the models including more runs using different data and different cuts of data, testing and training models on larger and more diverse datasets, including more EEG data from a more diverse range of participants and more diverse music selections, and analysis and modelling that take account of the effect of the individual's music preferences and brain signals on the EEG data. The genetic algorithm work could also potentially be used on larger EEG datasets to select relevant channels as well as relevant features.

References

Chen, T., Tang, K., Chen, G., Tang K., and Yao, X. 2012. *A Large Population Size Can be Unhelpful in Evolutionary Algorithms*. Theoretical Computer Science, vol. 436, pp. 54-70.

Fachner, J. and Stegemann, T. 2013. *Electroencephalography and Music Therapy: On the Same Wavelength?* Music and Medicine. July, pp. 1-6.

Fdez, J., Guttenbery, N., Witkowski, O., and Pasquali, A. 2021. *Cross-Subject EEG-Based Emotion Recognition Through Neural Networks with Stratified Normalization*. Frontiers in Neuroscience, vol. 15: 626277.

Nakisa, B., Rastgoo, M.N., Tjondronegoro, D., and Chandran, V. 2018. *Evolutionary Computation Algorithms for Feature Selection of EEG-Based Emotion Recognition Using Mobile Sensors*. Expert Systems with Applications, vol. 93, pp. 143-155.

Milne, L.K., Gedeon, T. and Skidmore, A.K., 1995. *Classifying Dry Sclerophyll Forest from Augmented Satellite Data: Comparing Neural Network, Decision Tree & Maximum Likelihood*. In Proceedings Australian Conference on Neural Networks.

Rahman, J.S., Gedeon, T., Caldwell, S. and Jones, R., 2020, July. *Brain Melody Informatics: Analysing Effects of Music on Brainwave Patterns*. In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8.

Sabeti, M., Boostani, R., Katebi, S.D., and Price, G.W. 2007. *Selection of Relevant Features for EEG Signal Classification of Schizophrenic Patients*. Biomedical Signal Processing and Control, vol. 2, pp. 122-134.

Thaut, M.H. 2015. *Chapter 8 – Music as therapy in early history*. Progress in Brain Research, vol. 217. pp. 143-158.

Von Ravenzwaaij, D. and Ioannidis, J.A.P. 2019. *True and False Positive Rates of Different Criteria if Evaluating Statistical Evidence from Clinical Trials*. BMC Medical Research Methodology, vol. 19: 218.

Wilkins, R.W., Hodges, D.A., Laurienti, P.J., Steen, M. and Burdette, J.H. 2014. *Network Science and the Effects of Music Preference on Functional Brain Connectivity: From Beethoven to Eminem*. Scientific Reports, vol. 4: 6130.