frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 20 June 2023
DOI 10.3389/frai.2023.1154663

OPEN ACCESS

EDITED BY
Ricky J. Sethi,
Fitchburg State University, United States

REVIEWED BY
Charles Courchaine,
National University, United States
F. Amilcar Cardoso,
University of Coimbra, Portugal

*CORRESPONDENCE
Paul J. Zak
✉ paul@neuroeconomicstudies.org

RECEIVED 03 February 2023
ACCEPTED 09 May 2023
PUBLISHED 20 June 2023

CITATION
Merritt SH, Gaffuri K and Zak PJ (2023) Accurately predicting hit songs using neurophysiology and machine learning. *Front. Artif. Intell.* 6:1154663. doi: 10.3389/frai.2023.1154663

Accurately predicting hit songs using neurophysiology and machine learning

Sean H. Merritt[1], Kevin Gaffuri[1] and Paul J. Zak[1,2]*

[1]Center for Neuroeconomics Studies, Claremont Graduate University, Claremont, CA, United States, [2]Immersion Neuroscience, Henderson, NV, United States

Identifying hit songs is notoriously difficult. Traditionally, song elements have been measured from large databases to identify the lyrical aspects of hits. We took a different methodological approach, measuring neurophysiologic responses to a set of songs provided by a streaming music service that identified hits and flops. We compared several statistical approaches to examine the predictive accuracy of each technique. A linear statistical model using two neural measures identified hits with 69% accuracy. Then, we created a synthetic set data and applied ensemble machine learning to capture inherent non-linearities in neural data. This model classified hit songs with 97% accuracy. Applying machine learning to the neural response to 1st min of songs accurately classified hits 82% of the time showing that the brain rapidly identifies hit music. Our results demonstrate that applying machine learning to neural data can substantially increase classification accuracy for difficult to predict market outcomes.

KEYWORDS

prediction, immersion, music, neurophysiology, classification

Introduction

Every day, 24,000 new songs are released worldwide (Pandora, 2018). That's 168,000 new songs every week. People are drowning in choices. The surfeit of choices makes it difficult for streaming services and radio stations to identify songs to add to playlists. Music distribution channels use both human listeners and artificial intelligence models to identify new music that is likely to become a hit. Unfortunately, the accuracy of predictions has generally been low (Prey, 2018). This has been called the "Hit Song Science" problem (McFee et al., 2012). The inability to predict hits means that artists are often underpaid for their work and music labels misallocate production and marketing resources when seeking to build audiences for new music (Byun, 2016). The inability to curate desirable music also causes audiences move between platforms searching for music they enjoy (Prey, 2018).

People want new music, but generally prefer songs similar to those they already know (Ward et al., 2014; Askin and Mauskapf, 2017). Music streaming services have invested in technologies to identify and introduce new music customized to subscribers' existing playlists. Spotify does this with "Discover Weekly," a playlist of 30 new songs subscribers receive every Monday morning. Pandora classifies new music using 450 attributes in its Music Genome Project and introduces new music using a service called "Personalized Soundtracks" (Carbone, 2021). Tracking what people add to their playlists boosts the likelihood of songs showing up in related playlists thereby building support leading to a hit (Turk, 2021). Nevertheless, less than 4% of new songs will become hits (Interiano et al., 2018).

Background

Predicting hits in entertainment is a long-standing problem (Litman, 1983). Predicting hit movies has been no better than a coin flip even after considering the director, the stars, budget, time of year of release, and whether or not the movie is a sequel (Chang and Ki, 2005; Sharda and Delen, 2006; Lash and Zhao, 2016). Various methods have been used to predict hit music, including the analysis of lyrics, blog postings, social media mentions, and brain activity (Dhanaraj and Logan, 2005; Abel et al., 2010; Berns and Moore, 2012; Singhi and Brown, 2014; Araujo et al., 2017). Yet, predictive accuracy for most studies is quite low.

The Hit Song Science problem is a subset of research known as Hit Science that seeks to predict whether entertainment content will be popular (Yang et al., 2017). This approach typically extracts components of content and applies machine learning to predict hits (Ni et al., 2011). When seeking to predict songs, this approach has typically used audio elements such as tempo, time signature, length, loudness, genre, lyric sentimentality, and instrument types (Herremans et al., 2014). Various classification techniques have been applied and yet predictive accuracy continues to be low (Raza and Nanath, 2020; Shambharkar et al., 2021).

Ex post, experts offer rationale for why hits were "inevitable" (Rodman, 2020). Yet, the apparent inevitability that some entertainment will become popular is challenged by studies that use ex-ante self-report for prediction (Morton, 1996). Direct and indirect measures of self-reported "liking" poorly predict aggregate outcomes (Hazlett and Hazlett, 1999; Wolfers and Zitzewitz, 2004; Bar-Anan et al., 2010; Cyders and Coskunpinar, 2011; John et al., 2017). When assessing music, "liking" is often anchored to familiarity resulting in poor ratings for unfamiliar songs (Ward et al., 2014). Moreover, using Likert self-report scales to predict popularity may be asking too much of study participants. Music is meant to elicit emotional responses that arise outside of conscious awareness and are often poorly reported (Thomas and Diener, 1990; Robinson and Clore, 2002). One way to avoid the inaccuracy of self-report is to directly measure neurophysiologic responses to music.

Music is an effective way to influence people's emotional states (Fitch, 2006). Music and language likely co-evolved, with the first evidence of musical instruments appearing in Paleolithic bone flutes 40,000 years ago (Conard et al., 2009). Structural features of music, including melody, tempo, key, and rhythm influence the emotions people experience (Scherer and Zentner, 2001). Lyrics, which add the human voice to music, generally increase emotional responses to music, especially for sad songs (Ali and Peynircioglu, 2006; Mori and Iwanaga, 2014). Songs are so effective at influencing emotions that they work at scale to build and sustain communities. Examples include Gregorian chants, the sacred music of Bach, the Catholic Tridentine Mass sung in Latin, Navajo priest singers, the singing of monks in the Buddhist and Daoist traditions, and Pygmy honey-gathering songs (Schippers, 2018). Songs are thought to be a way to socially regulate emotional responses, thereby influencing behavior (Hou et al., 2017). Contemporary music can be similarly powerful, influencing fads such as swing dancing, country line dancing, the Macarena, and Gangnam Style, as well as cadences used at military boot camps to teach recruits to march. Hit songs, due to their broad popularity, have an outsized influence on the emotional states of people, if only temporarily. The desire to experience an emotional state may explain why some songs become hits (Schulkind et al., 1999; Schellenberg and von Scheve, 2012).

Emotional responses emanant from multiple brain regions rather than being localized to a one or a few structures (Adolphs and Anderson, 2018). In addition to the auditory cortex, music has been shown to activate brain areas associated with processing emotions (amygdala, orbitofrontal cortex) and long-term memory retrival (hippocampus) (Koelsch et al., 2006; Levitin and Tirovolas, 2009). The multiple regions of the brain activated by music mean that peripheral rather than central measures of neural activity may better capture the response of neural circuits that process emotional stimuli (Mauss and Robinson, 2009) including responses to music (Coutinho and Cangelosi, 2011; Koelsch, 2018). This is consistent with the James-Lange theory of emotion in which neurophysiologic responses induce an emotional feeling (Derryberry and Tucker, 1992; McGaugh and Cahill, 2003; Barrett, 2006; Kreibig, 2010; Barrett and Westlin, 2021). While there is no one best way to measure neurophysiologic responses to emotional stimuli, peripheral measures appear to be more robust than central measures (Golland et al., 2014). For these reasons, the study here measured peripheral rather than central neurophysiologic responses.

State of the art

Our point of departure from the extant literature is to examine if neural measures can be used to predict hit music in order to address the Hit Song Science problem. While neural activity has been shown to add predictive power to self-reports, neural signals alone generally have poorly predictive accuracy for population outcomes (Falk et al., 2010, 2011; Berkman and Falk, 2013; Dmochowski et al., 2014; Genevsky et al., 2017). For example, a study using functional MRI to predict music popularity showed an improvement over self-report but predictive accuracy was still well-below 50% (Berns and Moore, 2012). One reason for this may be the inherent non-linearity of neural signals used as inputs into linear predictive models. While some researchers have directly modeled the nonlinear components of neural responses (Barraza et al., 2015), this is atypical.

A recent approach seeks to predict outcomes from neural data using machine learning. Machine learning more effectively integrates non-linear effects into predictions (Guixeres et al., 2017; Wei et al., 2018). Nevertheless, machine learning analyses may be subject to overfitting the data (Lemm et al., 2011). Overfitting can be reduced by using a limited the number of neural data streams (Jabbar and Khan, 2015), an approach we take here. Our analysis compares the classification accuracy of traditional linear predictive models to machine learning models using neurophysiologic measures alone.

Rather than choose a machine learning approach a priori (Raza and Nanath, 2020), we instead used an ensemble method (González et al., 2020). This approach estimates multiple individual models that are weighted and combined in order to increase predictive accuracy. Herein we apply an ensemble learning technique called bagging, also known as bootstrap aggregating, that avoids overfitting of data by reducing variance. Bagging also effectively

captures high-dimensional data (Zhang and Ma, 2012) and can be used used on weak learners with high variance and low bias (Alelyani, 2021).

Methods

In this section we detail the procedures through which the songs were chosen and market impact measures were obtained. We also describe the design of the laboratory experiment used to collect neurophysiologic data while participants listened to songs. While a commercial neurophysiology platform was used to capture neural responses, we derived two novel measures from these data that we anticipated would add insight into why some songs become hits. This section also outlines the data analytics methodology and provides a rationale for the machine learning approach we apply to predict hit songs from neurophysiologic measures.

Participants

Thirty-three participants (47% female) were recruited from the Claremont Colleges and surrounding community. Participants ranged in age from 18 to 57 (M = 24.25, SD = 10.47). This study was approved by the Institutional Review Board of Claremont Graduate University (#3574) and all participants gave written informed consent prior to inclusion. The data were anonymized by assigning an alphanumeric code to each participant.

Procedure

After consent, participants were seated and fitted with Rhythm + PPG cardiac sensors (Scosche Industries, Oxnard, CA). Music was played through a speaker system to groups of 5–8 participants in a medium-sized lab. Participants were informed that they would listen to 24 recent songs and asked about their preferences for each one. They then completed a short survey on demographics. The study lasted ∼1 h and participants were paid $15 for their time. Figure 1 shows the study timeline.

Neurophysiology

A commercial platform (Immersion Neuroscience, Henderson, NV) was used to measure neurophysiologic responses. Neurophysiologic immersion combines signals associated with attention and emotional resonance collected at 1 Hz. The attentional response is associated with dopamine binding to the prefrontal cortex while emotional resonance is related to oxytocin release from the brainstem (Barraza and Zak, 2009; Zak and Barraza, 2018; Zak, 2020). Together these neural signals accurately predict behaviors after a stimulus, especially those that elicit emotional responses (Lin et al., 2013; Barraza et al., 2015). The Immersion Neuroscience platform ingests device-agnostic heart rate data to infer neural states from activity of the cranial nerves using the downstream effects of dopamine and oxytocin (JeŽová et al., 1985; Zak, 2012; Barraza et al., 2015). The algorithms that measure immersion from cranial nerve activity are cloud-based and the platform provides an output file used in the analysis. We chose to measure neurologic immersion for this study because singing induces oxytocin release (Keeler et al., 2015) as does listening to music (Nilsson, 2009; Ooishi et al., 2017), though the effect is inconsistent (Harvey, 2020). The Immersion omnibus measure was expected to be more predictive than oxytocin alone or peripheral neural measures such as electrodermal activity (Ribeiro et al., 2019). Whether neurologic immersion can accurately classify hit songs is a new use of this measure.

The independent variables were average immersion for each song as well as two additional variables we derived from immersion data. The first we call peak immersion, defined as

$$\int_{t=0}^{T} (v_{it} > M_i) d_t / Im_i$$

where v_{it} is average neurophysiologic immersion across participants in song i at time t to the end of the song at time T, M_i is the median of the average time series of immersion for the duration of song i plus the standard deviation of song i across all participants who listened to that song divided by the sum of total immersion Im_i for song i. That is, peak immersion cumulates the highest immersion moments during the song relative to the song's total immersion. The second variable we created is called retreat. Neurologic retreat cumulates the lowest 20% of immersion averaged across participants for each song.

Songs

Staff from an online streaming service choose 24 songs for this study without input from the researchers. The streaming service also provided the definition of hits or flops. This resulted in a "clean" experiment as song choice could not be cherry-picked for the study and the criterion for a hit was established in advance. Thirteen songs were deemed "hits" with over 700,000 streaming listens, while the other 11 were flops. The songs had been released for no more than 6 months and spanned genres that included rock (Girl In Red "Bad Idea"), hip-hop (Roddy Rich "The Box"), and EDM (Tons and I "Dance Monkey"). Song order was counterbalanced and song start and stop times were synchronized with physiologic data. The 24 songs were used as the unit of analysis.

Surveys

After each song, participants were asked to rank how much they liked the song (1 to 10), if they would replay the song (0, 1), recommend the song to their friends (0, 1), if they had heard it previously to assess familiarity (0, 1), and if they found the song offensive (0, 1). We also showed participants lyrics from the song and lyrics created by the researchers and asked them to identify the song lyrics to measure their memory of the song (0, 1).

FIGURE 1
Timeline of the experiment.

Market data

The streaming service provided the researchers with market data from their platform. These included the number of song streams that varied from 4,000 (NLE Choppa "Dekario") to over 32 million ("Dance Money"). Additional data included the number of streaming stations that carried the song and online likes.

Statistical analysis

We used a sequence of statistical approaches, increasing in sophistication, to assess the predictive accuracy of neurophysiological variables. This was done so that the models can be directly compared. The analysis begins with tests of mean differences for self-report and neurophysiologic data comparing hits and flops using Student's *t*-tests (for readability, denoted "*t*-test"). Parametric relationships were examined using Pearson correlations while logistic regressions were estimated to establish predictive accuracy. Sensitivity analysis was conducted by analyzing the 1st min of data and re-assessing the likelihood of a song being a hit.

In order to improve predictive accuracy, we trained a bagged machine learning (ML) model. Bagged models are a type of ensemble ML model that tests several machine learning algorithms in an attempt to improve accuracy above that of a single model (Dietterich, 2000). Bagged models do this by taking the output of each model individually and making a prediction based on the weighted average prediction of each model. The SuperLearner package in R was used to train and test the weighted bagged models. We included common machine learning classification algorithms in the analysis, including logistic regression, k-nearest neighbors, neural nets, and support vector machines.

Logistic regression can be considered a machine learning method since it is designed as a statistical binary classifier. Support vector machines (SVMs) trains on data by fitting a hyperplane to separate classifications. These hyperplanes can be non-linear making them well-suited for neurophysiologic data. K-nearest neighbors (KNN) uses training data to create boundaries between different classification labels. It does this by iterating through each data point and using the k-nearest observations to determine boundaries for classification. Artificial neural networks (ANN) attempt to make predictions in a way that mimics the neural patterns of the brain. It takes each variable as an input and uses a series of linear and non-linear transformations to map them into outputs. These transformations are weighted using a backpropagation algorithm seeking to improve predictive accuracy (James et al., 2013).

Bagged ML maximizes predictive accuracy by comparing the predicted value of each algorithm using a training set to the actual value. It then combines algorithms by minimizing cross-validated risk (Van der Laan et al., 2007; Polley and van der Laan, 2010) weighting each one by its contribution to accuracy. The final predicted value is calculated as the sum of the predicted value for each algorithm multiplied by the derived weights,

$$\hat{Y}_i = \sum_{j=1}^{N} \beta_j \hat{Y}_{ij},$$

where β_j is the weight of algorithm j and $\hat{Y}_{ij}$ is the predicted value for song i by algorithm j. For details, see Polley and van der Laan (2010).

To find optimal parameter settings we used 5 fold cross-validation. The logistic regression used the optimal cost settings (1, 10, 100), the number of neighbors for KNN was (3, 5, 8, 10), cost (1, 10, 100) and kernel (radial, polynomial, hyperbolic tangent) were used for SVM, and an activation function (linear, softmax), while layer size (1, 5, 10), and decay (0, 1, 10) were used for the ANN. Optimal settings were identified as logit C = 1, KNN k = 3, SVM C = 10, kernel = tanh, and ANN function = softmax, layer size = 5, and decay = 1.

Small data sets are not appropriate for machine learning as they lead to high bias in their results (Vabalas et al., 2019). To address this, we created a synthetic set data with 10,000 observations using the synthpop package in R (Nowok et al., 2016). This standard automated procedure creates observations by repeatedly randomly sampling the joint distribution of the data. This technique is used when obtaining large datasets is infeasible, including analyses of computer vision (Mayer et al., 2018), sensitive information like hospital records (Tucker et al., 2020), and with unbalanced data (He et al., 2008; Luo et al., 2018). One-half of the synthetic data was used to train the bagged ML model and tune the hyperparameters. The other half of the synthetic data was used to test it. The Appendix compares the observed data to the synthetic dataset and discusses the methodology used to generate these data. Means, standard deviations and correlations were statistically identical. All participant data were used to train the models and to generate predictions.

Results

In the following subsections, we compare the predictive accuracy of self-reported song preferences to models using neural variables. Several measures of market impact are related to self-reports and neural variables in order to provide a baseline. Significant relationships are then used to build classifiers.

Neurophysiologic data are first used to estimate perhaps the simplest classifier, a logistic regression to establish a second baseline for ML model comparison. ML predictive accuracy is assessed using neural data from participants listening to complete songs as well as to data from the 1st min of each song. The latter is included to demonstrate the robustness of our results. Analyses of possible overfitting using K-fold cross-validation are also reported to establish model appropriateness. In is important to note that only two neurophysiologic measures served as the foundation for the ML models. This streamlines the interpretation of the findings and reduces the likelihood of overfitting.

Self-report

Self-reported liking was statistically related to the number of streams ($r = 0.54$, $N = 24$, $p = 0.002$) when analyzing participants who were familiar with the songs. Liking was not predictive for stations ($r = -0.08$, $N = 24$, $p = 0.701$) or online likes ($r = 0.38$, $N = 24$, $p = 0.060$). When analyzing songs that were unfamiliar to participants, the relationship between likes and online streams disappeared ($\beta = -0.13$, $N = 24$, $p = 0.387$). This suggests an endogeneity problem: are liked songs familiar or are more familiar songs liked? In order to avoid this issue, we analyzed data only from songs with which participants were unfamiliar. These data were aggregated to the song level and were used for all subsequent analyses.

For unfamiliar songs, self-reported liking was statistically identical for hits and flops ($M_{hit} = 4.49$, $M_{flop} = 4.48$; $t\ (22) = -00.05$, $p = 0.963$, $d = -00.02$). The same held for recommend [$t\ (22) = -00.21$, $p = 0.829$, $d = -00.09$], offensive [$t\ (22) = -00.44$, $p = 0.664$, $d = -00.18$] and lyrics [$t\ (20) = -00.58$, $p = 0.571$, $d = -00.25$]. None of the self-report measures correlated with steams ($r = 0.16$, $p = 0.46$), stations (-00.31, $p = 0.140$), or online likes ($r = 0.05$, $p = 0.980$).

Neurophysiologic responses

Hit songs had higher immersion than flops ($M_{hit} = 4.17$, $M_{flop} = 4.10$; $t\ (22) = -02.34$, $p = 0.028$, $d = -00.95$) while neurologic retreat and peak immersion did not differ between hits and flops (retreat: $t\ (22) = 2.01$, $p = 0.057$, $d = 0.82$; peak: $t\ (22) = -00.14$ $p = 0.887$, $d = -00.06$). Immersion was not correlated with streams ($r = 0.03$, p = 0.870), nor was peak immersion (r = 0.26, p = 0.202), while neurologic retreat trended toward significance ($r = -00.39$, $p = 0.057$). Immersion was negatively related to age ($r = -00.31$, $p < 0.001$) and varied by gender (Male: 4.06, Female: 4.20; $t\ (650) = -03.04$ $p < 0.001$, $d = -00.26$).

Accuracy

Logistic regression models were used to assess whether neurophysiologic measures could predict hits. Model 1 only included immersion. Then Model 2 added retreat to test if it would improve accuracy. Both Model 1 and Model 2 were significantly better at predicting hits than chance ($F = 5.48$, $p = 0.023$; $F = 3.27$, $p = 0.05$). Model 1 and Model 2 correctly classified hits and flops 66% and 70% of the time, respectively. Model 2 classified hits with 69% accuracy and flops with 62% accuracy using only the two neurophysiologic variables (ps $<$ 0.001). While retreat and immersion are correlated with each other ($r = -0.51$, $p = 0.011$), Model 2 did not suffer from multicollinearity (VIF = 1.07). The results were robust to the inclusion of an indicator for offensive lyrics ($p = 0.68$). As expected the neurophysiologic variables were not statistically related to the self-reported desire to replay the song, recommend the song to others, or the number of online likes for each song (ps > 0.68).

Machine learning

A bagged ML model was trained on one-half of the synthetic data using immersion and retreat as independent variables. The ML approaches that contribute more have a higher coefficient and lower risk (Polley and van der Laan, 2010). The k-nearest neighbor model contributed the most (coef = 0.98, risk = 0.018) followed by a neural net (coef = 0.012, risk = 0.18). A logistic regression and support vector machines contributed nothing to an accurate classification.

The bagged ML model was able to accurately classify the type of song 97.2% of the time. This statistically greater than the base rate (Successes = 4,800, $N = 5{,}000$, $p < 0.001$) using the exact binomial test (Hollander and Wolfe, 1973). Examining specificity and sensitivity, hits were classified correctly 96.6% of the time and flops were classified with 97.6% accuracy (Figure 2).

Next, we assessed the bagged ML model's ability to predict hits from the original 24 song data set. The bagged ML model accurately classified songs with 95.8% which is significantly better than the baseline 54% frequency (Success = 23, $N = 24$, $p < 0.001$). Only one song, Evil Spider, was classified incorrectly. This song was a flop with nearly 54,000 streams but was classified as a hit due to its high immersion.

We conducted a bootstrap procedure with 1,000 iterations on both the bagged ML model and the logistic model to compare their accuracy for hits and flops. The logit was trained on one data set ($N = 5{,}000$) and then assessed for accuracy on another set of data ($N = 5{,}000$) for each iteration. The bagged ML model predicted hits ML: CI = [1, 1]; Logistic: CI = [0.67, 0.73]; $t\ (1{,}998) = -115.86$, $p < .001$) and flops (ML: CI = [0.82,1.00]; Logistic: CI = [0.59, 0.63]; $t\ (1{,}998) = -121.13$, $p < 0.001$) better than the logistic model.

The model was assessed for overfitting by running a 10-fold cross validation on the bagged ML and comparing the predictive accuracy of the training set, test set, and observed data. This analysis shows that the bagged ML does not appear to overfit the test data as the accuracy is high and consistent (James et al., 2013). As expected, the accuracy is higher on the training and test synthetic data across the k-folds (∼0.99) compared to the $N = 24$ observed data (∼0.96). Nevertheless, across the three data sets the accuracy of the model is high, similar, and consistent (Figure 3).

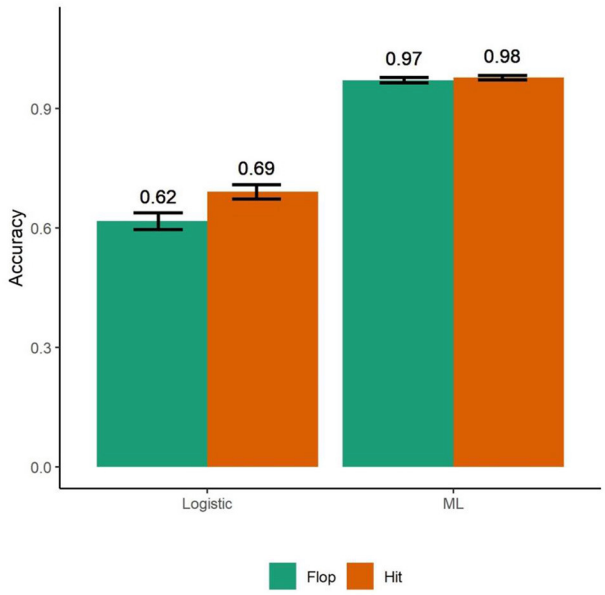

FIGURE 2
A traditional logistic regression model using neurophysiologic immersion and retreat as independent variables correctly classified hits with 69% accuracy and flops with 62% accuracy ($N = 5{,}000$). A bagged machine learning model using the same two independent variables had accuracy of 96.6% in classifying hits while flops were classified with 97.6% accuracy. Bars are standard deviations.

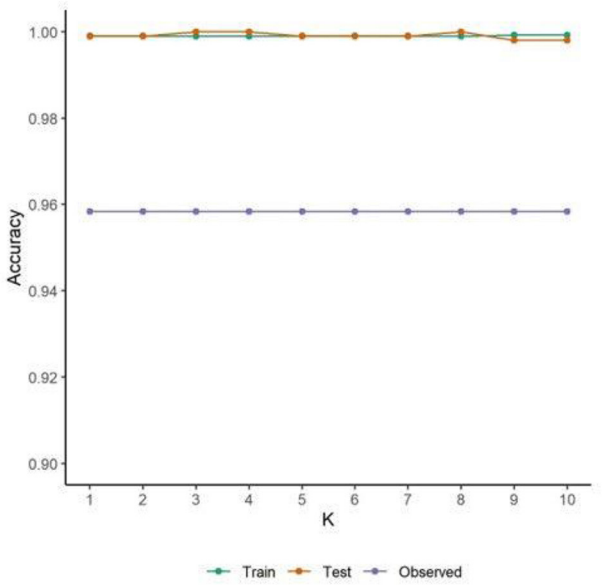

FIGURE 3
A 10-fold cross validation indicates the bagged ML model consistently predicts test data with ~99% accuracy. Accuracy for the observed data is consistent at ~96% indicating that the model does not overfit the data.

1 min of data

To establish the robustness and practical applications of our findings, we analyzed the accuracy of neurophysiology collected from the first 1 min of data to identify hits. We ran the same logistic and bagged ML models described previously using only the data from the 1st min of each song. We did not find a significant relationship between immersion (OR = 362.25, $N = 24$, $p = 0.101$) or retreat (OR = 27175.63, $N = 24$, $p = 0.406$) and hit songs. However, using immersion and retreat, we were able to correctly classify songs 66% of the time using a logistic regression. Specificity and sensitivity were moderate at 77% and 56%, respectively.

We also created another synthetic data set to train a bagged machine learning model. Our bagged ML model had overall accuracy of 74%. It predicted hit songs with 82% accuracy and flops with 66% accuracy (Figure 4). Using the bagged ML model on the original data, we found that it was able to predict hits and flops 66% of the time. Bootstrapping the results, the bagged ML model outperformed the logistic model in classifying hit songs (ML: CI = [0.80, 0.82], Logistic: CI = [0.75, 0.7848]; $N = 5{,}000$, t (1,998) = −41.76, $p < 0.001$), and flops (ML: CI = [0.65, 0.70], Logistic: CI = [0.54, 0.58]; t (1,998) = −22.61, $p < 0.001$).

Discussion

The key contribution of the present study is to demonstrate that neurophysiologic measures accurately identify hit songs while self-reported "liking" is unpredictive. In addition, we showed that neurophysiology, combined with machine learning, substantially improves the classification of hit songs when compared to linear statistical models. Our goal was to provide a methodology that other researchers can use to predict hit songs of different genres, in different geographic locations, and for study populations with different demographics. The approach described here should also be tested for its ability to accurately predict hits for other forms of entertainment that are known to be difficult to ascertain, including movies, TV shows, and social media posts in order to confirm and extend our results. Indeed, our use of a commercial neuroscience platform makes such extensions feasible for non-neuroscientists.

Forecasting one's own behavior based on reflection is fraught and using self-report to predict market outcomes of entertainment is nearly always a fool's errand (Sheeran, 2002; Woodside and Wilson, 2002; Brenner and DeLamater, 2016). Even experts cannot identify high quality goods and services from their imitators (Ashenfelter and Jones, 2013; Almenberg et al., 2014). While people want to hear new music, they prefer music that is similar to familiar songs generating a bias in self-reports (Ward et al., 2014). The Hit Song Science problem is typically addressed by mining very large datasets (McFee et al., 2012). We took a different approach, collecting neurophysiologic data from a moderate number of people to predict aggregate outcomes. We showed that self-reported liking only identifies hit songs if one was already familiar with the song. This is most likely due to endogeneity: participants are more likely to report they like a song if they have heard it often. Once we removed participants' familiar songs, self-reported liking ceased to predict hits.

The use of neurophysiologic data to predict aggregate outcomes is an approach that has been labeled "brain as predictor" or "neuroforecasting." This approach captures neural activity from

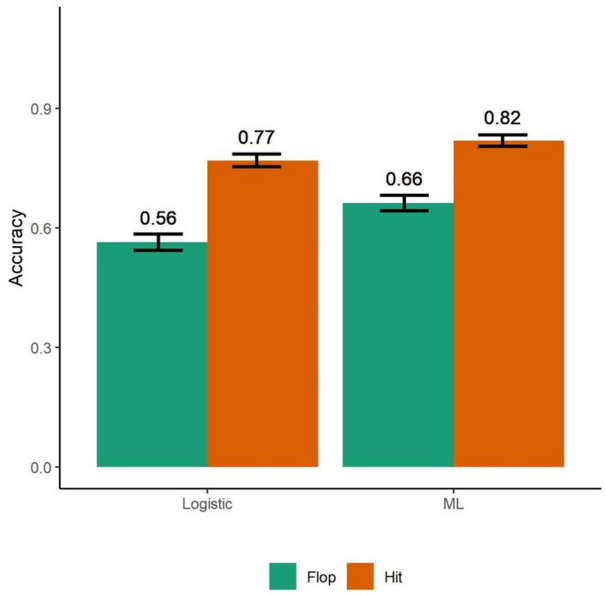

FIGURE 4
A logistic regression model trained on neurophysiologic responses for the 1st min of songs correctly classified hits with 77% accuracy and flops with 56% accuracy ($N = 10{,}000$). The bagged machine learning model classified 74% of songs correctly. Hits were classified with 82% accuracy while flops were identified accurately 66% of the time. Model training used half of the synthetic data set ($N = 5{,}000$) using a bootstrapped evaluation of 5,000 observations per iteration for 1,000 iterations. Bars are standard deviations.

a small group of participants to predict population outcomes (Berkman and Falk, 2013; Dmochowski et al., 2014; Genevsky et al., 2017). Neurologic data have been shown to predict outcomes more accurately than self-reports for sunscreen use, smoking reduction strategies, watching TV, and crowdfunding requests (Falk et al., 2010, 2011; Yang et al., 2015; Genevsky et al., 2017; Zak, 2022). While estimating predictive models using neural data is an improvement over poorly-predicting self-reported measures, the accuracy of neural forecasts have generally been no better than 50%. The closest published study to the report here used fMRI data from 28 people to predict the popularity of 20 songs. One brain region, the nucleus accumbens, was correlated with aggregate outcomes but was only able to correctly classify hits with 30% accuracy (Berns and Moore, 2012).

Our analysis showed that two measures of neurophysiologic immersion in music identified hits and flops with 69% accuracy using a traditional linear logistic regression model. A logistic regression using only the 1st min of the song was nearly as accurate at 66% and was 77% accurate at classifying which songs were hits. This is a substantial improvement over the existing literature. Most of the models cited above using neural data to predict aggregate outcomes have focused on attentional responses. The neurophysiologic data we used convolves attentional and emotional responses and this may account for our improved predictive accuracy (Lench et al., 2011; Zak and Barraza, 2018; Zak, 2020). Emotional responses are a key component of persuasive communication because emotions capture the subjective value of an experience (Barraza et al., 2015; Cacioppo et al., 2018; Falk and Scholz, 2018; Doré et al., 2019). The analysis here indicates that emotional responses also appear to determine which songs become hits.

Applying a bagged machine learning model to neural data improved its predictive accuracy from 69% to 97%. We also demonstrated the robustness and practical use of our approach by correctly classifying hits with 82% accuracy using the 1st min of songs. It is worth noting that no demographic or self-report data were used in these models. Further, our findings are unlikely due to chance. Machine learning using neural data has been used to identify mental illness (Stahl et al., 2012; Khodayari-Rostamabad et al., 2013; Amorim et al., 2019), epilepsy (Shoeb and Guttag, 2010; Buettner et al., 2019), stress (Subhani et al., 2017), and to recognize emotions (Zhang et al., 2020). Market researchers have applied machine learning to neural data to predict views of Superbowl ads and behavioral responses to advertising (Guixeres et al., 2017; Wei et al., 2018). As of this writing, machine learning models of music have used lyrical content rather than neural data to classify hits with only moderate accuracy (Dhanaraj and Logan, 2005; Singhi and Brown, 2014). We extended these approaches by using neural responses to music from a modest number of people to identify hits. Future work could connect neural responses to lyric classifications for additional insights.

Rather than choose a single machine learning algorithm, our use of an ensemble model eliminated a manual search for the best approach. The analysis showed that a k-nearest neighbors' (KNN) algorithm was responsible for the majority of the explanatory power. While machine learning models have been called "black boxes," our analysis showed that hit songs have higher immersion than flops and do so with a large effect size (*Cohen's d = 0.95*). Hits also produced less neurologic retreat than flops with a similar effect size (*Cohen's d = 0.82*). Another reason to use machine learning to classify hits is that neurophysiologic data are inherently non-linear. Unlike logistic regressions, KNN's incorporate non-linear relationships making it ideally suited to neural data.

It is noteworthy that only three neurophysiologic variables were used in the analysis. Among these, only two of the variables, immersion and retreat, were statistically associated with classifying hits. The results are consistent with the intuition that hits are expected to have higher immersion and produce less retreat than do flops. The parsimony of models with two variables, and their associated high accuracy, supports the measurement of peripheral neural measures for this application. While the classification accuracy using linear relationships of these variables, as in the logistic regression we reported, was only moderate, the ML approach utilized non-linear relationships which are more appropriate for neural responses (Timme and Lapish, 2018).

While the accuracy of the present study was quite high, there are several limitations that should be addressed in future research. First, our sample was relatively small so we are unable to assess if our findings generalize to larger song databases. The large effect sizes indicate the results are likely to be similarly accurate if other songs were tested. We also created a synthetic data set to train the machine learning model. These data, while generated from human neural responses, may have overweighted subtle relationships not

evident in the original data. Nevertheless, this approach has become standard when access to large samples, such as in experimental studies as reported here, is not available (Hoffmann et al., 2019). The use of synthetic data allows researchers to gather less direct participant data with a small or no loss in accuracy. While we found high accuracy using the observed data, we did not have access to an outside sample of songs to validate the model further. This means our model might have overfitted the data.

Conclusion

Measuring emotional responses using neuroscience technologies provides a new way for artists, record producers, and streaming services to delight listeners with new music. Our contribution is to show that omnibus neuroscience measurements from the peripheral nervous system quite accurately classify hits and flops. Rather than asking users if they "like" a new song, wearable neural technologies, like those in this study, could assess the neural value of content automatically. Steaming services' "Discover Weekly," and "Personalized Soundtracks" could more effectively build playlists of desired new music by measuring neurologic immersion when users listen to just the 1st min of a new song. Music from users' existing playlists could also be chosen by using neurologic immersion to identify mood states as was recently shown (Merritt et al., 2022). Our findings, if replicated, indicate that as neuroscience technologies enter into general use, the ability to curate music and other forms of entertainment to give people just what they want will improve existing recommendation engines benefiting artists, distributors, and consumers.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://doi.org/10.3886/E152561V1.

Ethics statement

The studies involving human participants were reviewed and approved by IRB Claremont Graduate University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PZ designed the study. SM and KG performed the statistical analysis. SM, KG, and PZ wrote the manuscript. All authors discussed the results and provided critical feedback and helped shape the research, analysis and manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1154663/full#supplementary-material

References

Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., and Siehndel, P. (2010). "Analyzing the blogosphere for predicting the success of music and movie products," in *2010 International Conference on Advances in Social Networks Analysis and Mining* (pp. 276-280). IEEE. doi: 10.1109/ASONAM.2010.50

Adolphs, R., and Anderson, D. J. (2018). *The Neuroscience of Emotion: A New Synthesis*. Berlin: Princeton University Press. doi: 10.23943/9781400889914

Alelyani, S. (2021). Stable bagging feature selection on medical data. *J. Big Data*, 8, 1–18. doi: 10.1186/s40537-020-00385-8

Ali, S. O., and Peynircioglu, Z. F. (2006). Songs and emotions: are lyrics and melodies equal partners? *Psychol. Music*, 34, 511–534. doi: 10.1177/0305735606067168

Almenberg, J., Dreber, A., and Goldstein, R. (2014). *Hide the Label, Hide the Difference?* New York, NY: American Association of Wine Economists.

Amorim, E., Van der Stoel, M., Nagaraj, S. B., Ghassemi, M. M., Jing, J., O'Reilly, U. M., et al. (2019). Quantitative EEG reactivity and machine learning for prognostication in hypoxic-ischemic brain injury. *Clin. Neurophysiol.* 130, 1908–1916. doi: 10.1016/j.clinph.2019.07.014

Araujo, C. V., Neto, R. M., Nakamura, F. G., and Nakamura, E. F. (2017). "Predicting music success based on users' comments on online social networks," in *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web* (pp. 149-156). doi: 10.1145/3126858.3126885

Ashenfelter, O., and Jones, G. V. (2013). The demand for expert opinion: Bordeaux wine. *J. Wine Econ.* 8, 285–293. doi: 10.1017/jwe.2013.22

Askin, N., and Mauskapf, M. (2017). What makes popular culture popular? *Product features and optimal differentiation in music. Am. Sociol. Rev.*, 82, 910–944. doi: 10.1177/0003122417728662

Bar-Anan, Y., Wilson, T. D., and Hassin, R. R. (2010). Inaccurate self-knowledge formation as a result of automatic behavior. *J. Exp. Soc. Psychol.*, 46, 884–894. doi: 10.1016/j.jesp.2010.07.007

Barraza, J., and Zak, P. (2009). Empathy toward strangers triggers oxytocin release and subsequent generosity. *Annals New York Acad. Sci.* 1167, 182–189. doi: 10.1111/j.1749-6632.2009.04504.x

Barraza, J. A., Alexander, V., Beavin, L. E., Terris, E. T., and Zak, P. J. (2015). The heart of the story: peripheral physiology during narrative exposure predicts charitable giving. *Biol. Psychol.* 105, 138–143. doi: 10.1016/j.biopsycho.2015.01.008

Barrett, L. F. (2006). Are emotions natural kinds?. *Perspect. Psychol Sci* 1, 28–58. doi: 10.1111/j.1745-6916.2006.00003.x

Barrett, L. F., and Westlin, C. (2021). "Navigating the science of emotion," in *Emotion Measurement* (pp. 39-84). Woodhead Publishing. doi: 10.1016/B978-0-12-821124-3.00002-8

Berkman, E. T., and Falk, E. B. (2013). Beyond brain mapping: using neural measures to predict real-world outcomes. *Curr. Dir. Psychol. Sci.*, 22, 45–50. doi: 10.1177/0963721412469394

Berns, G. S., and Moore, S. E. (2012). A neural predictor of cultural popularity. *J. Cons. Psychol.* 22, 154–160. doi: 10.1016/j.jcps.2011.05.001

Brenner, P. S., and DeLamater, J. (2016). Lies, damned lies, and survey self-reports? *Identity as a cause of measurement bias. Soc. Psychol. Quart.* 79, 333–354. doi: 10.1177/0190272516628298

Buettner, R., Frick, J., and Rieg, T. (2019). "High-performance detection of epilepsy in seizure-free EEG recordings: A novel machine learning approach using very specific epileptic EEG sub-bands," in *ICIS*. Munich.

Byun, C. (2016). *The Economics of the Popular Music Industry: Modelling from Microeconomic Theory and Industrial Organization*. Berlin: Springer.

Cacioppo, J. T., Cacioppo, S., and Petty, R. E. (2018). The neuroscience of persuasion: a review with an emphasis on issues and opportunities. *Soc. Neurosci.* 13, 129–172. doi: 10.1080/17470919.2016.1273851

Carbone, S. (2021). Spotify vs Pandora. What's in the box? SoundGuys. Avilable onlinw at: https://www.soundguys.com/spotify-vs-pandora-36915/ (accessed April 15, 2021).

Chang, B. H., and Ki, E. J. (2005). Devising a practical model for predicting theatrical movie success: focusing on the experience good property. *J. Media Econ.* 18, 247–269. doi: 10.1207/s15327736me1804_2

Conard, N. J., Malina, M., and Münzel, S. C. (2009). New flutes document the earliest musical tradition in southwestern Germany. *Nature*, 460, 737–740. doi: 10.1038/nature08169

Coutinho, E., and Cangelosi, A. (2011). Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, 11, 921. doi: 10.1037/a0024700

Cyders, M. A., and Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: is there overlap in nomothetic span and construct representation for impulsivity?. *Clin. Psychol. Rev.*, 31, 965–982. doi: 10.1016/j.cpr.2011.06.001

Derryberry, D., and Tucker, D. M. (1992). Neural mechanisms of emotion. *J. Consult. Clin. Psychol.* 60, 329. doi: 10.1037/0022-006X.60.3.329

Dhanaraj, R., and Logan, B. (2005). "Automatic Prediction of Hit Songs," in *ISMIR* (pp. 488-491).

Dietterich, T. G. (2000). "Ensemble methods in machine learning," in *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1* (Berlin; Heidelberg: Springer), 1–15.

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., Parra, L. C., et al. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nat. Commun.* 5, 1–9. doi: 10.1038/ncomms5567

Doré, B. P., Tompson, S. H., O'Donnell, M. B., An, L. C., Strecher, V., Falk, E. B., et al. (2019). Neural mechanisms of emotion regulation moderate the predictive value of affective and value-related brain responses to persuasive messages. *J. Neurosci.* 39, 1293–1300. doi: 10.1523/JNEUROSCI.1651-18.2018

Falk, E., and Scholz, C. (2018). Persuasion, influence, and value: perspectives from communication and social neuroscience. *Annu. Rev. Psychol.* 69, 329–356. doi: 10.1146/annurev-psych-122216-011821

Falk, E. B., Berkman, E. T., Mann, T., Harrison, B., and Lieberman, M. D. (2010). Predicting persuasion-induced behavior change from the brain. *J. Neurosci* 30, 8421–8424. doi: 10.1523/JNEUROSCI.0063-10.2010

Falk, E. B., Berkman, E. T., Whalen, D., and Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health.* 30, 177. doi: 10.1037/a0022259

Fitch, W. T. (2006). The biology and evolution of music: a comparative perspective. *Cognition* 100, 173–215. doi: 10.1016/j.cognition.2005.11.009

Genevsky, A., Yoon, C., and Knutson, B. (2017). When brain beats behavior: neuroforecasting crowdfunding outcomes. *J. Neurosci.* 37, 8625–8634. doi: 10.1523/JNEUROSCI.1633-16.2017

Golland, Y., Keissar, K., and Levit-Binnun, N. (2014). Studying the dynamics of autonomic activity during emotional experience. *Psychophysiology* 51, 1101–1111. doi: 10.1111/psyp.12261

González, S., García, S., Del Ser, J., Rokach, L., and Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. *Inform. Fusion*, 64, 205–237. doi: 10.1016/j.inffus.2020.07.007

Guixeres, J., Bigné, E., Ausin Azofra, J. M., Alcañiz Raya, M., Colomer Granero, A., Fuentes Hurtado, F., et al. (2017). Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. *Front. Psychol.* 8, 1808. doi: 10.3389/fpsyg.2017.01808

Harvey, A. R. (2020). Links between the neurobiology of oxytocin and human musicality. *Front. Hum. Neurosci.* 14, 350. doi: 10.3389/fnhum.2020.00350

Hazlett, R. L., and Hazlett, S. Y. (1999). Emotional response to television commercials: facial EMG vs. self-report. *J. Adv. Res.* 39, 7–7.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE International Joint Conference on Neural Networks*. IEEE World Congress on Computational Intelligence (IEEE), 1322–1328.

Herremans, D., Martens, D., and Sörensen, K. (2014). Dance hit song prediction. *J. New Music Res.* 43, 291–302. doi: 10.1080/09298215.2014.881888

Hoffmann, J., Bar-Sinai, Y., Lee, L. M., Andrejevic, J., Mishra, S., Rubinstein, S. M., et al. (2019). Machine learning in a data-limited regime: augmenting experiments with synthetic data uncovers order in crumpled sheets. *Sci. Adv.* 5, eaau6792. doi: 10.1126/sciadv.aau6792

Hollander, M., and Wolfe, D. A. (1973). *Non-parametric Statistical Methods*. New York: John Wiley and Sons. pp. 15–22.

Hou, J., Song, B., Chen, A. C., Sun, C., Zhou, J., Zhu, H., et al. (2017). Review on neural correlates of emotion regulation and music: implications for emotion dysregulation. *Front. Psychol.* 8, 501. doi: 10.3389/fpsyg.2017.00501

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., Komarova, N. L., et al. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *R. Soc. Open Sci.* 5, 171274. doi: 10.1098/rsos.171274

Jabbar, H., and Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Comp. Sci. Commun. Instrument. Dev.* 3, 163–172. doi: 10.3850/978-981-09-5247-1_017

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer (112, 18). doi: 10.1007/978-1-4614-7138-7

JeŽová, D., Jurčovičová, J., Vigaš, M., Murgaš, K., and Labrie, F. (1985). Increase in plasma ACTH after dopaminergic stimulation in rats. *Psychopharmacology* 85, 201–203. doi: 10.1007/BF00428414

John, L. K., Emrich, O., Gupta, S., and Norton, M. I. (2017). Does "liking" lead to loving? The impact of joining a brand's social network on marketing outcomes. *J. Market. Res.* 54, 144–155. doi: 10.1509/jmr.14.0237

Keeler, J. R., Roth, E. A., Neuser, B. L., Spitsbergen, J. M., Waters, D. J. M., Vianney, J. M., et al. (2015). The neurochemistry and social flow of singing: bonding and oxytocin. *Front. Hum. Neurosci.* 3, 518. doi: 10.3389/fnhum.2015.00518

Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G. M., Bruin, de., and MacCrimmon, H. (2013). A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clin. Neurophysiol.* 124, 1975–1985. doi: 10.1016/j.clinph.2013.04.010

Koelsch, S. (2018). Investigating the neural encoding of emotion with music. *Neuron* 98, 1075–1079. doi: 10.1016/j.neuron.2018.04.029

Koelsch, S., Fritz, T., Cramon, D. Y., et al. (2006). Investigating emotion with music: an fMRI study. *Hum. Brain Mapp.* 27, 239–250. doi: 10.1002/hbm.20180

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biol. Psychol.* 84, 394–421. doi: 10.1016/j.biopsycho.2010.03.010

Lash, M. T., and Zhao, K. (2016). Early predictions of movie success: the who, what, and when of profitability. *J. Manag. Inform. Sys.* 33, 874–903. doi: 10.1080/07421222.2016.1243969

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K. R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Lench, H. C., Flores, S. A., and Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. *Psychol. Bull.* 137, 834. doi: 10.1037/a0024244

Levitin, D. J., and Tirovolas, A. K. (2009). Current advances in the cognitive neuroscience of music. *Ann. N. Y. Acad. Sci.* 1156, 211–231. doi: 10.1111/j.1749-6632.2009.04417.x

Lin, P-. Y., Grewal, N. S., Morin, C., Johnson, W. D., and Zak, P. J. (2013). Oxytocin increases the influence of public service advertisements. *PLoS ONE* 8, 6934. doi: 10.1371/journal.pone.0056934

Litman, B. R. (1983). Predicting success of theatrical movies: an empirical study. *J. Pop. Cult.* 16, 159. doi: 10.1111/j.0022-3840.1983.1604_159.x

Luo, D., Zou, Y., and Huang, D. (2018). "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Interspeech*. p. 152–156.

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., et al. (2018). What makes good synthetic training data for learning disparity and optical flow estimation?. *Int. J. Comput. Vision.* 126, 942–960.

McFee, B., Bertin-Mahieux, T., Ellis, D. P., and Lanckriet, G. R. (2012). "The million song dataset challenge," in *Proceedings of the 21st International Conference on World Wide Web* (pp. 909-916). doi: 10.1145/2187980.2188222

McGaugh, J. L., and Cahill, L. (2003). "Emotion and memory: Central and peripheral contributions," in *Series in Affective Science. Handbook of Affective Sciences*, eds R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Eds.), (p. 93–116). Oxford University Press.

Merritt, S. H., Krouse, M., Alogaily, R. S., and Zak, P. J. (2022). Continuous neurophysiologic data accurately predict mood and energy in the elderly. *Brain Sci.* 12, 1240. doi: 10.3390/brainsci12091240

Mori, K., and Iwanaga, M. (2014). Pleasure generated by sadness: effect of sad lyrics on the emotions induced by happy music. *Psychol. Music*, 42, 643–652. doi: 10.1177/0305735613483667

Morton, A. (1996). Folk psychology is not a predictive device. *Mind* 105, 119–137. doi: 10.1093/mind/105.417.119

Ni, Y., Santos-Rodriguez, R., Mcvicar, M., and Bie, D. e. T. (2011). "Hit song science once again a science," in 4th *International Workshop on Machine Learning and Music*.

Nilsson, U. (2009). Soothing music can increase oxytocin levels during bed rest after open-heart surgery: a randomised control trial. *J. Clin. Nurs.* 18, 2153–2161. doi: 10.1111/j.1365-2702.2008.02718.x

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *J. Statist. Softw.* 74, 1–26.

Ooishi, Y., Mukai, H., Watanabe, K., Kawato, S., and Kashino, M. (2017). Increase in salivary oxytocin and decrease in salivary cortisol after listening to relaxing slow-tempo and exciting fast-tempo music. *PLoS ONE*, 12, e0189075. doi: 10.1371/journal.pone.0189075

Pandora. (2018). *Personal Communication to PJZ*, New York City.

Polley, E. C., and van der Laan. (2010). M. J. Super learner In prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 266. Available online at: https://biostats.bepress.com/ucbbiostat/paper266 (accessed April 15, 2023).

Prey, R. (2018). Nothing personal: algorithmic individuation on music streaming platforms. *Media Cult. Soc.* 40, 1086–1100. doi: 10.1177/01634437177 45147

Raza, A. H., and Nanath, K. (2020). "Predicting a hit song with machine learning: is there an apriori secret formula?" in 2020 *International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)* (pp. 111-116). IEEE. doi: 10.1109/DATABIA50434.2020.9190613

Ribeiro, F. S., Santos, F. H., Albuquerque, P. B., and Oliveira-Silva, P. (2019). Emotional induction through music: measuring cardiac and electrodermal responses of emotional states and their persistence. *Front. Psychol.* 10, 451. doi: 10.3389/fpsyg.2019.00451

Robinson, M. D., and Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychol. Bull.* 128, 934. doi: 10.1037/0033-2909.128.6.934

Rodman, S. (2020). David Foster on his hit songs and working with Whitney Houston, Celine Dion, and more in his new Netflix documentary. yahoo! news, Available online at: https://news.yahoo.com/david-foster-hit-songs-working-134058286.html (accessed May 6, 2021).

Schellenberg, E. G., and von Scheve, C. (2012). Emotional cues in American popular music: five decades of the Top 40. *Psychol. Aesthet. Creat. Arts*, 6, 196. doi: 10.1037/a0028024

Scherer, K. R., and Zentner, M. R. (2001). "Emotional effects of music: production rules," in *Music and emotion: theory and research*, eds Juslin, P. N. and Sloboda, J. A. (Oxford: Oxford University Press) 5, 361–387.

Schippers, H. (2018). Community music contexts, dynamics, and sustainability. *The Oxford Handbook of Community Music*, 23-42. doi: 10.1093/oxfordhb/9780190219505.013.29

Schulkind, M. D., Hennis, L. K., and Rubin, D. C. (1999). Music, emotion, and autobiographical memory: they're playing your song. *Memory Cogn.* 27, 948–955. doi: 10.3758/BF03201225

Shambharkar, P. G., Singh, A., and Yadav, A. (2021). Hit song prediction: using ant colony optimization for feature selection. *New Arch-Int. J. Contemp. Architect.* 8, 1298–1305.

Sharda, R., and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.*, 30, 243–254. doi: 10.1016/j.eswa.2005.07.018

Sheeran, P. (2002). Intention—Behavior relations: a conceptual and empirical review. *Eur. Rev. Soc. Psychol.* 12, 1–36. doi: 10.1080/14792772143000003

Shoeb, A. H., and Guttag, J. V. (2010). "Application of machine learning to epileptic seizure detection," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 975-982).

Singhi, A., and Brown, D. G. (2014). Hit Song Detection Using Lyric Features Alone. *Proceedings of International Society for Music Information Retrieval*.

Stahl, D., Pickles, A., Elsabbagh, M., Johnson, M. H., and Team, B. A. S. I. S. (2012). Novel machine learning methods for ERP analysis: a validation from research on infants at risk for autism. *Dev. Neuropsychol.* 37, 274–298. doi: 10.1080/87565641.2011.650808

Subhani, A. R., Mumtaz, W., Saad, M. N. B. M., Kamel, N., and Malik, A. S. (2017). Machine learning framework for the detection of mental stress at multiple levels. *IEEE Access* 5, 13545–13556. doi: 10.1109/ACCESS.2017.2723622

Thomas, D. L., and Diener, E. (1990). Memory accuracy in the recall of emotions. *J. Pers. Soc. Psychol.* 59, 291. doi: 10.1037/0022-3514.59.2.291

Timme, N. M., and Lapish, C. (2018). A tutorial for information theory in neuroscience. *eNeuro* 5, 18. doi: 10.1523/ENEURO.0052-18.2018

Tucker, A., Wang, Z., Rotalinti, Y., and Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit. Med.* 3, 1–13.

Turk, V. (2021). How to bust your Spotify feedback loop and find new music. *Wired*. Available online at: https://www.wired.com/story/spotify-feedback-loop-find-new-music/ (accessed April 15, 2021).

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14, e0224365. doi: 10.1371/journal.pone.0224365

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* 6, 1309. doi: 10.2202/1544-6115.1309

Ward, M. K., Goodman, J. K., and Irwin, J. R. (2014). The same old song: the power of familiarity in music choice. *Mark. Lett.* 25, 1–11. doi: 10.1007/s11002-013-9238-1

Wei, Z., Wu, C., Wang, X., Supratak, A., Wang, P., Guo, Y., et al. (2018). Using support vector machine on EEG for advertisement impact assessment. *Front. Neurosci.* 12, 76. doi: 10.3389/fnins.2018.00076

Wolfers, J., and Zitzewitz, E. (2004). Prediction markets. *J. Econ. Perspect.* 18, 107–126. doi: 10.1257/0895330041371321

Woodside, A. G., and Wilson, E. J. (2002). Respondent inaccuracy. *J. Advert. Res.*, 42, 7–18. doi: 10.2501/JAR-42-5-7-18

Yang, L. C., Chou, S. Y., Liu, J. Y., Yang, Y. H., and Chen, Y. A. (2017). "Revisiting the problem of audio-based hit song prediction using convolutional neural networks," in 2017 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 621-625). IEEE. doi: 10.1109/ICASSP.2017.7952230

Yang, T., Lee, D. Y., Kwak, Y., Choi, J., Kim, C., Kim, S. P., et al. (2015). Evaluation of TV commercials using neurophysiological responses. *J. Physiol. Anthropol.*, 34, 1–11. doi: 10.1186/s40101-015-0056-4

Zak, P. J. (2012). *The Moral* Molecule: *The Source of Love and Prosperity*. New York, NY: Random House.

Zak, P. J. (2020). Neurological correlates allow us to predict human behavior. *The Scientist. Oct.* 1, 3.

Zak, P. J. (2022). *Immersion: The Science of the Extraordinary and the Source of Happiness*. NY: Lioncrest.

Zak, P. J., and Barraza, J. A. (2018). Measuring immersion in Experiences with biosensors. Proceedings of the 11th *International Joint Conference on Biomedical Engineering Systems and Technologies*. doi: 10.5220/0006758203030307

Zhang, C., and Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. *Springer Sci. Bus. Media*. 3, 7. doi: 10.1007/978-1-4419-9326-7

Zhang, J., Yin, Z., Chen, P., and Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. *Inform. Fusion* 59, 103–126. doi: 10.1016/j.inffus.2020.01.011