OCTOBER 24 2025

Explainable machine learning models for classifying reactions within crowd noise during men's collegiate basketball games

Mitchell C. Cutler ; Jason Bickmore; Mark K. Transtrum ; Katrina Pedersen ; Shannon Proksch ; Eli Farrer; Kent Gee

J. Acoust. Soc. Am. 158, 3456–3471 (2025)
https://doi.org/10.1121/10.0039709

JASA ARTICLE

Explainable machine learning models for classifying reactions within crowd noise during men's collegiate basketball games

Mitchell C. Cutler,[1] Jason Bickmore,[1] Mark K. Transtrum,[1] Katrina Pedersen,[1] Shannon Proksch,[2,3] Eli Farrer,[1] and Kent Gee[1,a)]

[1]*Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA*

[2]*Department of Psychology, Augustana University, Sioux Falls, South Dakota 57197, USA*

[3]*Program in Neuroscience, Center for Interdisciplinary Studies, Augustana University, Sioux Falls, South Dakota 57197, USA*

ABSTRACT:

Crowds at collegiate basketball games react acoustically to events on the court in many ways, including applauding, chanting, cheering, and making distracting noises. Acoustic features can be extracted from recordings of crowds at basketball games to train machine learning models to classify crowd reactions. Such models may help identify crowd mood, which could help players secure fair contracts, venues refine fan experience, and safety personnel improve emergency response services or to minimize conflict in policing. By exposing the key features in these models, feature selection highlights physical insights about crowd noise, reduces computational costs, and often improves model performance. Feature selection is performed using random forests and least absolute shrinkage and selection operator logistic regression to identify the most useful acoustic features for identifying and classifying crowd reactions. The importance of including short-term feature temporal histories in the feature vector is also evaluated. Features related to specific 1/3-octave band shapes, sound level, and tonality are highly relevant for classifying crowd reactions. Additionally, the inclusion of feature temporal histories can increase classifier accuracies by up to 12%. Interestingly, some features are better predictors of future crowd reactions than current reactions. Reduced feature sets are human-interpretable on a case-by-case basis for the crowd reactions they predict.

 https://doi.org/10.1121/10.0039709

(Received 7 November 2024; revised 9 October 2025; accepted 9 October 2025; published online 24 October 2025)

[Editor: James F. Lynch] Pages: 3456–3471

I. INTRODUCTION

Collegiate men's basketball games have a complex mixture of sound sources, including the public address (PA) system, music from live bands, sounds from the officiating crew, and various unified sounds from the crowd in response to events on the court, referred to as "crowd reactions." Classifying crowd reactions may provide means to quantify the entertainment value added by individual athletes[1] (e.g., interest, attendance, etc.) through direct, real-time crowd engagement at sporting event venues, as opposed to indirect engagements through social media.[2,3] In light of recent developments in name, image, and likeness contracts in intercollegiate sports,[4–6] there may be many stakeholders[7] interested in this quantitative approach. Identifying which acoustic features are most useful for classifying acoustic crowd reactions and why they are important may also lead to insights into the collective behavior of crowds and be of interest to social psychology and the cognitive sciences.[8,9] Across other disciplines, acoustic monitoring of crowd reactions may help identify sentiment and mood changes[10] of a crowd and improve the ability to advise emergency response teams[11] or to minimize conflict in policing.[12]

Previous analyses of crowd noise have studied synchronization of voices and sound levels,[13,14] the effects of crowd noise on athletes,[15,16] identification of key moments in sporting events,[17] unsupervised classifications of acoustic spectra,[18–20] and supervised classification of general sentiment of crowds using neural networks and spectrograms.[10]

In this paper, machine learning models are built on the features described in Ref. 18 to classify crowd reactions at men's basketball games into one of four classes: *applause*, *chant*, *cheer*, or *distraction noise*. These classes are further defined in Ref. 19. The classification model developed in this study is composed of several binary classification models. The first one separates instances of sounds produced by the crowd (*crowd noise*) from instances without crowd sounds present (*non-crowd noise*), which may include sounds produced by individuals, music, the PA system, and the officiating crew. Instances of *crowd noise* may also include non-crowd noise (e.g., a buzzer going off while a crowd cheers). Other classifiers then separate anything labeled or classified as *crowd noise* into the four classes of reactions (i.e., there is one binary classifier for each class). Feature selection is performed via several methods to reduce the feature sets for all classifiers. This paper investigates why some features are more useful than others for classifying crowd reactions and examines and explains the relevance of the temporal histories of the most informative

a)Email: kentgee@byu.edu

https://doi.org/10.1121/10.0039709

features. Feature temporal histories increase in importance for some feature sets, and a few features relating to tonality, specific spectral shapes, and intensity have clear connections to physical phenomena and are useful in classifying several crowd reactions in the noisy setting of a basketball game.

A. Motivation for feature selection

Feature selection is used to improve model interpretability,[21] minimize the curse of dimensionality,[22] and reduce computational and data requirements by eliminating uninformative features. In some cases, it can also improve model performance. Unlike feature extraction methods, which perform dimension-reducing transformations on the feature set, feature selection methods choose a subset of the original features, which preserves the features' interpretability.[23] Ideally, the reduced feature set should contain a small number of features that are informative, interpretable, and relatively independent of each other.[24] Information provided by independent acoustic features is measured in several ways, such as by calculating each feature's correlation with crowd reactions or considering the accuracy of machine learning models that use these features to classify crowd reactions. Because the meaning of each feature remains the same during the process, feature selection is closely related to explainable artificial intelligence (AI),[25] which attempts to find the underlying rules behind complicated models. In this paper, various feature selection techniques are used to shed light on which features are useful for classifying acoustic crowd reactions. These techniques are mean decrease in accuracy (MDA), the Gini feature importance metric, the Gini feature importance metric with two kinds of penalties, and the least absolute shrinkage and selection operator (LASSO). Where possible, the selected features are linked to the physical characteristics of the crowd noise.

II. METHODS

In this section, acoustic data collection and preparation methods are explained, as well as machine learning feature generation. Correlations between features and the significance of these correlations in this project are also explained. We then introduce the model and its hyperparameters, review the five methods of feature selection used in this paper, and finally outline our approach to feature selection and hyperparameter tuning. Figure 1 provides a high-level sketch of the entire process, including data processing, feature generation, and feature selection.

A. Dataset

The 17.21 h of recordings used in this study came from ten Brigham Young University (BYU) men's basketball home games during the 2017–2018 and 2018–2019 seasons, with audiences ranging from 10 179 to 16 456 attendees, as reported by the venue.[26] All games were played in the same venue. Although reverberation times may have varied with attendance, reverberation time was not considered in this study. Each recording was made with a single microphone placed far enough from the crowd that no individual voice dominated. Pressure was sampled at 25 kHz or higher with a 24-bit system and a type 1, 12.7 mm diameter free-field microphone.[19] The average 1/3-octave band spectra of all ten games shared the same shape and were within 10 dB at all frequencies. Despite differences in attendance, the unweighted half-second equivalent continuous sound level ($Leq_{0.5\,s}$) distributions were similar for all games, with an L_{50} exceedance level of 91.7 dB.[18] The data were split by game into training, validation, and holdout test sets. While splitting the ten games into these datasets was not strictly random, games were divided with no pattern other than achieving desirable proportions (60-20-20 split). Details about these partitions in the data can be seen in Table I. Because the machine learning models created in this study will be applied to games outside of the dataset presented here, each game used in this study was assigned to one dataset (training, testing, or validation) to give a more robust and conservative estimate of the model's transferability. Because the crowd is the same throughout each game, splitting the data within a game might lead to false increases in model improvement. Validation data were used to tune model hyperparameters, and testing data were used to measure model accuracy during feature selection.

Seven human labelers listened to game recordings and manually labeled ten crowd reactions: singing, silence, cheer, positive chant, negative chant, applause, distraction noise, angry noise, disappointment, and surprise. These labelers met and agreed on the meaning of each of the ten labels. Each whole game was labeled by a single labeler. Some labelers labeled multiple games. To label a game, the labeler listened to the game audio and recorded the timestamps at which each reaction started and stopped in a simple labeling interface or a *.csv* file. Some games were labeled with the help of game footage to supplement the audio. While labelers met periodically to ensure labels were applied consistently, the consistency of human labeling was not systematically checked. The labeling process and seven of the ten crowd reaction labels are described in more detail in Ref. 19. This paper focuses on four classes of reactions: *applause, chanting* (positive chant and negative chant combined), *cheering*, and *distraction noise*. Additionally, instances where any crowd reaction label is active are separated from those where no label is active by the labels *any reaction* and *no reaction*, respectively. Human labelers noted other crowd reactions, such as *angry noise* and *singing*, but these labels are only included within the crowd reaction class of *any reaction* and not as their own classes because these reactions occur less frequently. Some crowd reactions overlap frequently, such as *applause* and *cheer*. Most crowd reactions do not have definite starting or stopping times, so some ambiguities are present in the labeling of the games, especially at the beginnings and ends of reactions. All unlabeled data are presumed not to contain any sort of crowd reaction. The percentage of data

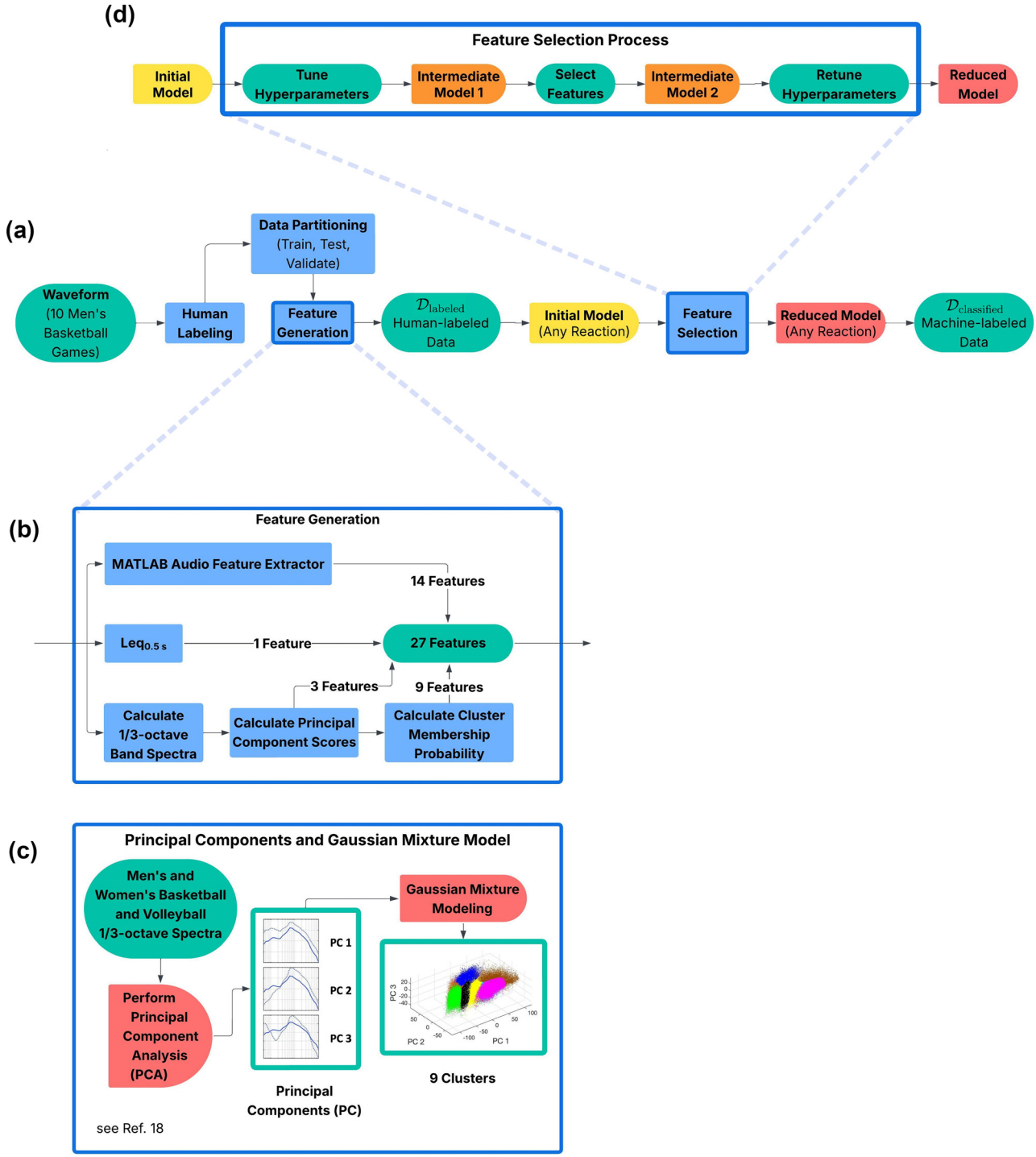

FIG. 1. (a) Overview of process from waveform to reduced model for *any reaction* classifier. (b) Feature generation process. (c) Creating the three principal components and nine clusters used in feature generation. (d) Feature selection process, including hyperparameter tuning, feature selection, and retuning. Hyperparameters were tuned with the validation data, and feature selection was performed with the testing data.

TABLE I. Splitting of the data into training, validation, and testing sets.

Set	No. of games	% of data	Total length (h)
Training	6	57.14	9.84
Validation	2	21.90	3.77
Testing	2	20.95	3.61

labeled as each reaction can be seen in Table II. There are differences in the proportion of classes in the training, validation, and testing sets because game events differ between games, as well as the makeup and behavior of the crowd (see Table II). This presents an opportunity to collect more data in the future to balance out the training, testing, and validation datasets.

https://doi.org/10.1121/10.0039709

TABLE II. Percentage of data, total length, and number of segments labeled as each reaction. A segment is a continuous portion of game audio.

Reaction	% of data, total length (h), no. of segments[a]		
	Training	Validation	Testing
Applause	5.97, 0.59, 394	2.17, 0.08, 56	1.33, 0.05, 41
Chant	5.85, 0.58, 265	4.19, 0.16, 86	4.50, 0.16, 89
Cheer	8.11, 0.8, 464	7.56, 0.29, 174	6.90, 0.25, 151
Distraction noise	10.2, 1.01, 365	8.62, 0.33, 128	6.79, 0.25, 105
Any reaction	31.8, 3.12, 1247	27.1, 1.02, 470	20.7, 0.75, 390
Unlabeled	68.2, 6.71, N/A	72.9, 2.75, N/A	79.3, 2.86, N/A

[a]N/A, not applicable.

B. Feature generation

The 27 features that are examined in this study were processed on half-second intervals. Coarser temporal resolutions cannot reasonably capture subtle dynamic changes in crowd reactions, but finer temporal resolutions were empirically found to not contain additional spectral information for 1/3-octave bands.[18,19] Fourteen of the features come from MATLAB's[27] Audio Feature Extractor,[28,29] using a linearly spaced power spectrum from 50 to 10 000 Hz with 2 Hz bins. Pitch[30] was not included because it is only valid when the acoustic signal is harmonic. Some features are transformed to make the units reflect logarithmic scaling in frequency and power. Features with values on a logarithmic-like scale, but with positive and negative values, are processed using hyperbolic arcsines. The extracted features with their transformation and brief descriptions are given in Table III (see Refs. 31–36).

Another feature, the unweighted half-second equivalent continuous sound level, or $\mathrm{Leq}_{0.5\,\mathrm{s}}$, is calculated directly from the audio pressure waveform. The remaining 12 features come from Ref. 18, including the three principal components shown in Fig. 1(c) and described in Table IV (referred to as PC 1, PC 2, and PC 3). These principal components come from a principal component analysis[37] performed on half-second intervals of 1/3-octave spectral bands [dB(z) re 20 μPa] ranging from 50 to 10 000 Hz from a dataset of 30 men's and women's intercollegiate basketball and volleyball games. Because the data are 1/3-octave band spectra, these principal components are referred to as principal spectral shapes. Although 24 of these principal spectral shapes were found in Ref. 18, only the first three are used here since they account for 87.5% of the variance in the dataset from which they were calculated. The first of these principal spectral shapes is comparable to $\mathrm{Leq}_{0.5\,\mathrm{s}}$. The second corresponds to mid-range frequencies, peaking just under 1 kHz, corresponding closely to sounds produced primarily by the crowd. The third corresponds to the spectral peakedness. Linear combinations of the three principal spectral shapes define a space. The data represented in this space were clustered into nine clusters using a Gaussian mixture model[37] in Ref. 18. These clusters are shown in Fig. 1(c). The probability of a point x in this three-dimensional space belonging to cluster $C = i$ given N clusters with weights w_i is

$$P(C = i|x) = \frac{f_X(x; \mu_i, \Sigma_i) w_i}{\sum_{j=1}^{N} f_X(x; \mu_j, \Sigma_j) w_j}, \tag{1}$$

where $f_X(x; \mu_i, \Sigma_i)$ is the Gaussian probability density function of the cluster i with mean μ_i and covariance Σ_i evaluated at the point x. These cluster membership probabilities, derived from Ref. 18, constitute the remaining nine features. The clusters were labeled by colors in Ref. 18 (green, pink, yellow, red, black, orange, cyan, blue, and brown), and for consistency, the same color labels refer to the same nine clusters in this paper. A brief description of each cluster is given in Table IV. These colors do not correspond to acoustic noise types (such as brown, pink, or white noise). It is important to note that data leakage is possible because the dataset used to calculate the principal spectral shapes included men's basketball games, which are also part of the training, validation, and testing data used in this study. The

TABLE III. Transformations and brief descriptions of features from MATLAB's Audio Feature Extractor toolbox (Ref. 28).

Feature	Transformation	Description
Spectral centroid (Ref. 31) (Centroid)	Log_{10}	First statistical moment
Spectral crest (Ref. 31) (Crest)	Log_{10}	Peakedness
Spectral decrease (Ref. 31) (Decrease)	Arcsinh $(x/0.006)$	Slope
Spectral entropy (Ref. 32) (Entropy)	None	Information entropy of the spectrum
Spectral flatness (Ref. 33) (Flatness)	$10\ \log_{10}$	Peakedness
Spectral flux (Ref. 34) (Flux)	$10\ \log_{10}(x/2 \times 10^{-5})$	Change in spectrum
Spectral kurtosis (Ref. 31) (Kurtosis)	Log_{10}	Fourth statistical moment
Spectral roll-off point (Ref. 34) (Roll-off)	Log_{10}	Frequency bounding 95% of energy
Spectral skewness (Ref. 31) (Skewness)	Arcsinh $(x/6)$	Third statistical moment
Spectral slope (Ref. 35) (Slope)	$\mathrm{Log}_{10}(-x)$	Slope
Spectral spread (Ref. 31) (Spread)	Log_{10}	Second statistical moment
Harmonic ratio (Ref. 36) (Harmonic)	None	Ratio of harmonic to total energy
Short time energy (STE)	$10\ \log_{10}(x/2 \times 10^{-5})$	Energy in signal
Zero-crossing rate (Cross rate)	None	Rate of signal crossing 0 Pa

TABLE IV. Descriptions of the three principal spectral shapes and nine Gaussian clusters. Cluster descriptions come from Ref. 18.

Feature	Description
PC 1	Principal spectral shape comparable to $Leq_{0.5\ s}$
PC 2	Principal spectral shape resembling crowd-produced sounds
PC 3	Principal spectral shape corresponding to spectral peakedness
Green cluster	Minimal noise
Pink cluster	Music
Yellow cluster	PA/individual noise
Red cluster	PA/music
Black cluster	Individual noise
Orange cluster	Moderate crowd noise
Cyan cluster	Music/moderate crowd noise
Blue cluster	High crowd noise
Brown cluster	Music/high crowd noise

process for generating each of these 27 features is summarized in Fig. 1(b).

C. Feature correlations

Because useful features are ideally strongly correlated with reaction classes while being relatively independent of each other, we first examine the correlation matrices of the features with each other and with the crowd reactions (see Fig. 2) before starting feature selection. This is done with the Pearson correlation coefficients in Fig. 2, which only show linear correlations and do not account for how features and reactions relate to each other temporally or nonlinearly. This also provides a way to validate later results.

Figure 2 suggests that PC 1, spectral flux, spectral slope, short time energy, and the $Leq_{0.5\ s}$ are all good predictors for distinguishing between *cheer* and *other reactions*. However, these particular features would not make a good feature subset together to classify *cheer* because they are strongly correlated with each other and are therefore redundant. A feature subset with just one of those features and another less strongly correlated feature such as spectral entropy or PC 2 would likely have better predictive properties, even though PC 2 and spectral entropy are less correlated with *cheer*.

D. Model and hyperparameter tuning

Random forests[40,41] are commonly used for feature selection.[42–47] They are fast to train, make robust predictions, can handle large numbers of features, and work naturally with the Gini feature importance metric. This study used scikit-learn's implementation of random forests.[48] Random forests have many hyperparameters, but Ref. 49 demonstrated that *max_features* (the number of features considered as candidate splitting features at each split) and *max_samples* (the number of data points each tree can train on, sampled with replacement) are two of the most important hyperparameters to tune. Reference 50 showed theoretically that larger forests improve performance and stabilize feature selection metrics; in practice, as many trees should be used as is computationally feasible.

A new hyperparameter, *time_steps_before*, is introduced to account for the time-dependent nature of the data. It takes on the values 0, 1, 2, 3, … and represents how many additional half-seconds of temporal history are used to predict the reaction at the current half-second time step. Some of the information that is useful for classifying crowd noise is contained in the temporal history of the acoustic features, not just in the features themselves at single points in time. During the feature selection process for the random forest models, feature importance scores were averaged over all half-second intervals.

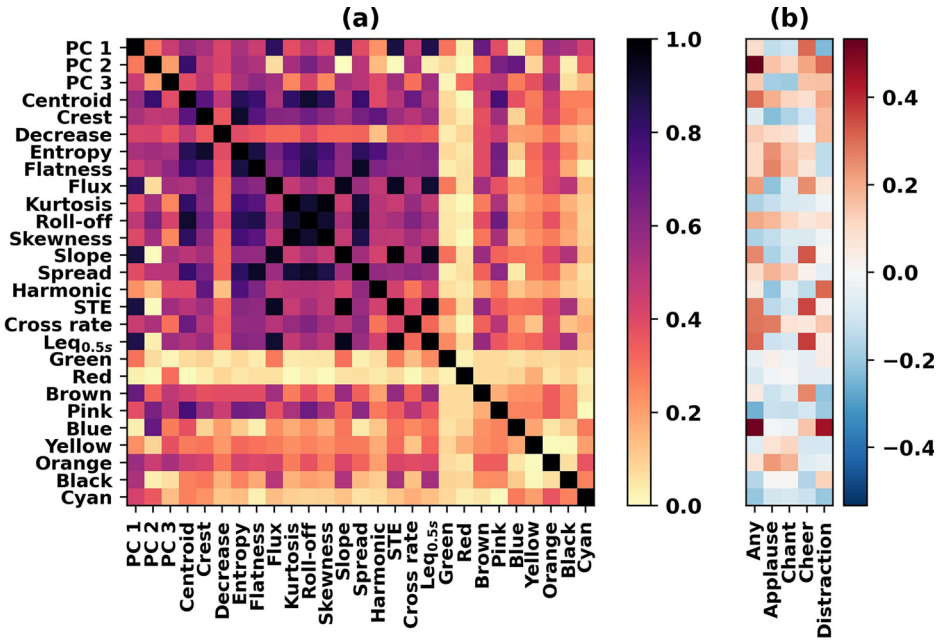

FIG. 2. Panel (a) shows the absolute value of the Pearson correlation coefficients between each pair of features for the training data (Refs. 38 and 39). Panel (b) shows the (signed) Pearson correlation coefficient between the features and each of the reactions studied in this paper for the training data (Refs. 38 and 39). Note that the correlations for individual crowd reactions include only the data labeled or classified as *any reaction*.

https://doi.org/10.1121/10.0039709

The *max_features* hyperparameter defaults to the square root of the number of features in scikit-learn. Because a smaller *max_features* encourages more diverse trees, stabilizes rankings for weak features, and allows moderately important features to be chosen more frequently,[49] *max_features* was set to $\log_2$. A grid search was performed to optimize *time_steps_before* (between 0 and 25 half-seconds with half-second increments) and *max_samples* (between 0.5 and 1.0 of the total number of data points with increments of 0.02), using forests of 1000 trees before beginning each feature selection process.

The cost surfaces were flat, varying by less than 4% in accuracy for all reactions, about half of which was from going from *time_steps_before* = 0 to *time_steps_before* = 2. Because the surfaces were so flat and a lower *max_samples* value promotes diverse trees, the grid search space was reduced to the values of *max_samples* between 0.60 and 0.80, with a 0.02 spacing, and the optimal values were chosen. Optimal values for *time_steps_before* for each reaction were 2 for *any reaction*, 12 for *applause*, 2 for *chant*, 3 for *cheer*, and 4 for *distraction noise*. For *applause*, the cost steadily improved for longer temporal histories (increased *time_steps_before*), while the costs for the other three reactions plateaued or decreased slightly. After optimizing over *time_steps_before* and *max_samples*, the random forests were tested with different forest sizes from 500 to 10 000 in increments of 500. The accuracy of the model varied little (less than 0.5%) between 3000 and 10 000 trees in every case, so 3000 trees were used to lower computational costs.

E. Feature selection methods

Five methods for feature selection are considered here. Four methods are examples of greedy algorithms[51] (i.e., features are removed one at a time based on which has the least predictive impact according to some metric). The fifth method is based on LASSO regression and is not a greedy algorithm.

The first greedy feature selection method uses MDA to determine feature importance. MDA is calculated by first training a model (with accuracy f_0) on a set of N features. Additional models are then trained on every possible subset of features of size $N-1$ (denote their accuracies by f_i, $i \in \{1, 2, \ldots, N\}$, where i is the index of the missing feature). The feature i that minimizes the mean decrease in accuracy (given by $f_0 - f_i$) is then dropped, and the process is repeated. In the most common form of MDA, only one model is trained at each step and the model is validated on the data with values of the i-th feature randomly permuted.[40] Although this form is much cheaper computationally, it has been criticized by Ref. 45 because it often forces the model to extrapolate, so it is avoided here, and instead, a separate model is trained on each feature subset before choosing which feature to drop.

The other three greedy feature selection methods in this paper use variations on the Gini feature importance metric, which is commonly used for both building random forests and measuring feature importance. References 40, 41, 47, and 52 describe the Gini feature importance metric and its computation. Although widely used, the Gini importance metric has two well-known flaws. The first is that high-cardinality features are often favored over low-cardinality features.[53,54] This is not an issue in this study because all the features are continuous variables. The second is that while it performs well at choosing informative features, it often favors correlated (i.e., redundant) features.[46] To address this, a variation of the Gini feature importance metric was proposed by Ref. 42 that explicitly penalizes correlated features. In this variation, for each feature x_i, the most correlated feature, x_{corr}, is identified. If the Gini importance of x_i is greater than that of x_{corr}, the feature importance is unchanged. Otherwise, the Gini importance of x_i is reduced by a factor of $(1 - \text{corr}_{\max})$, where $\text{corr}_{\max}$ is the correlation between x_i and x_{corr}. In this study, two correlation metrics are used: (1) the Pearson correlation coefficient, which measures linear correlations; and (2) the Spearman correlation coefficient, which measures rank-based correlations.[38,39] These three variants on the Gini feature importance metric will be referred to as "Gini" (no correlation penalty), "Gini–Pearson," and "Gini–Spearman" feature importance.

The fifth feature selection method is LASSO regression using a logistic regression model class.[55] In this method, the parameters associated with each feature in a logistic regression model are progressively pushed to zero as the strength of a one-norm regularization penalty term is increased. In this paper, the regularization penalty was increased by increments of 0.5. The features whose parameters remain non-zero the longest are considered the most useful for making predictions. Unlike the other four feature selection methods, LASSO regression is not a greedy method, and as such, eliminated features may reappear as the selection process progresses.

F. Feature selection pipeline

Figure 1(d) illustrates the feature selection process. First, a random forest model was trained on the whole dataset to distinguish between crowd reactions and non-crowd reactions ("Initial model"). The hyperparameters *time_steps_before* and *max_samples* were then tuned. These hyperparameters are discussed in detail in Sec. II D. This random forest model was then feature-reduced until model accuracy decreased significantly. (The resulting model had two features, the $\text{Leq}_{0.5\,\text{s}}$ and PC 2, and is described further in Sec. III A.) The hyperparameters *time_steps_before* and *max_samples* were then retuned. This reduced model was run on all available data $\mathcal{D}$, resulting in a subset, $\mathcal{D}_{\text{classified}}$, consisting of data that were classified by the model as containing a reaction. Another subset of $\mathcal{D}$ was $\mathcal{D}_{\text{labeled}}$, which consisted of data labeled as *any reaction* (i.e., data were labeled as either *applause*, *chant*, *cheer*, *distraction noise*, or *other reactions*). The set of data $\mathcal{D}_{\text{all reactions}} = \mathcal{D}_{\text{classified}} \cup \mathcal{D}_{\text{labeled}}$ was used for training, testing, and validating binary classifiers (random forest or LASSO regression) that distinguished *applause*, *chant*, *cheer*, or *distraction noise* from all other data in $\mathcal{D}_{\text{all reactions}}$, as indicated by Fig. 3(a).

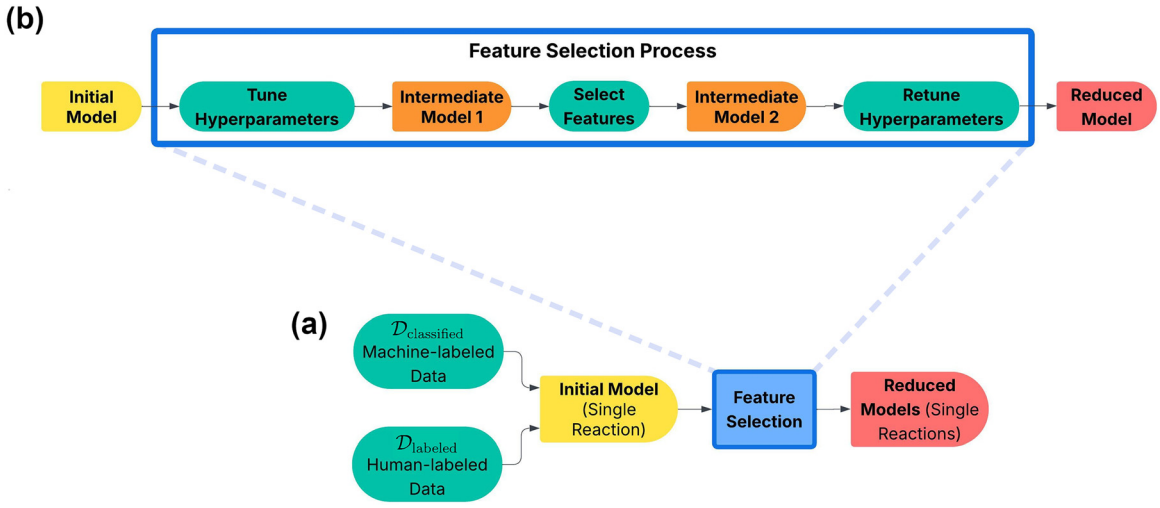

FIG. 3. (a) Process for obtaining reduced models for each single reaction classifier (e.g., *applause*, *cheer*, etc.). The output of the *any reaction* reduced model, $\mathcal{D}_{\text{classified}}$, is combined with human-labeled $\mathcal{D}_{\text{labeled}}$ to train the initial model for each of the five single reaction classifiers. (b) Feature selection process, identical to that in Fig. 1.

The hyperparameters *time_steps_before* and *max_samples* were tuned once for each of the classifiers used to distinguish between reactions in $\mathcal{D}_{\text{all reactions}}$ before beginning feature selection (see Sec. II D). Hyperparameters were never retuned during feature selection. After feature selection was performed on the models to distinguish between reactions, the hyperparameter *time_steps_before* was retuned to compare the optimal hyperparameter values before and after feature selection.

For every classification task in this pipeline (to create $\mathcal{D}_{\text{classified}}$ as well as to separate out individual reactions from $\mathcal{D}_{\text{all reactions}}$), several procedures were performed. First, larger classes were undersampled randomly to create balanced training, validation, and testing sets. Second, the random forest hyperparameters were tuned once for the full feature set (i.e., before feature selection).

As suggested by its name, random forest performance is slightly affected by which random seed is used.[44,56] Although these random fluctuations in performance are generally small and decrease with larger forest sizes, multiple random seeds were used in the tree-building algorithm at each step in the feature selection process to investigate the consistency of the results. For the three Gini-based methods, 50 random seeds were used at each step in the feature selection process before dropping the feature with the lowest average importance. The same was done with the MDA method, but with only 20 random seeds since the computational cost of feature selection was so much higher. LASSO methods involve a convex optimization problem, which has a unique solution and does not depend on random seed.

III. RESULTS

A. Detecting any crowd reactions

The accuracies of the *any reaction* binary classifiers as a function of the number of features remaining are shown in Fig. 4, while the top five features are given in order in Table V. The top five features are those that are cut from the model last during the feature selection process. Features were removed one at a time moving left to right in Fig. 4, so the right side of Fig. 4 illustrates these top features' impact on accuracy as each is cut. This figure shows that the *any reaction* classifier only requires information from a few top features before the classification accuracy begins to decrease significantly. This could be because other features contain irrelevant or redundant information, such as spectral crest, which is irrelevant, or PC 1, which is redundant with $\text{Leq}_{0.5\,\text{s}}$ (see Fig. 2). If they contain redundant information, then other subsets of features could be chosen with similar predictive power.

PC 2 was consistently the highest-ranked feature across all feature selection methods, except LASSO regression, which chose it as the second-most-important feature. PC 2 was found in Ref. 18 to distinguish between data points with more high- or low-frequency content in a dataset of men's and women's basketball and volleyball games. Figure 5(b) shows the effect of PC 2 on the average 1/3-octave band spectrum. The $\text{Leq}_{0.5\,\text{s}}$ was the second-highest-ranked feature in four out of the five feature selection methods. Taken together, *any reaction* generally corresponds to points that have both more high- than low-frequency content and are acoustically intense, as seen in Fig. 5(a), which shows how the density of PC 2 and $\text{Leq}_{0.5\,\text{s}}$ varies for *any reaction* (red) and *no reaction* (black).

As stated in Sec. II F, a reduced feature model separated crowd noise (crowd reactions) from other sounds before proceeding to the next part of the study. Only the two features PC 2 and$\text{Leq}_{0.5\,\text{s}}$ were used in constructing this model.

B. Distinguishing specific reactions

The two most useful features (PC 2 and the $\text{Leq}_{0.5\,\text{s}}$) for identifying *any reaction* were used to train and tune a random forest classifier. This classifier was used to create the $\mathcal{D}_{\text{all reactions}}$ dataset described in Sec. II F. These data were

https://doi.org/10.1121/10.0039709

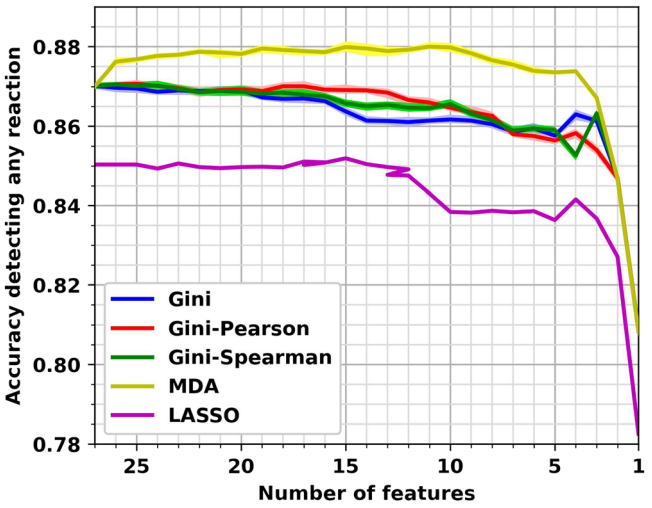

FIG. 4. Accuracy of the *any reaction* classifier models on the testing data as the number of features decreases. The two highest-ranked features in every random forest-based model were PC 2 and the $Leq_{0.5\ s}$. The shaded area around each line indicates the minimum and maximum accuracies across various random seeds, while the bold line indicates the mean. The minimum and maximum accuracies are difficult to see because they are close to the mean.

then used for training classifiers to distinguish individual crowd reactions from other crowd reactions within the *any reaction* dataset. Feature selection was performed on these classifiers and all features. A summary of the top features and optimized hyperparameters chosen by the feature selection methods is given in Table VI. More details are provided in Sec. IV.

1. Optimal number of features and feature rankings

The accuracy of the binary classifiers for specific crowd reactions as a function of number of features is shown in Fig. 6. Similar to the *any reaction* classifier, the number of features can be significantly reduced without changing the model accuracies. Transparent bands around the accuracy curves show the minimum and maximum accuracies after running the feature selection process 20–50 times with different random seeds. The LASSO regression method is deterministic, so its accuracy does not vary with the random seed, but increasing the regularization strength can cause it to reselect features that it had previously dropped, which is why the number of features sometimes increases.

Although the exact number of features to which the dataset can be reduced is somewhat subjective and depends on the feature selection method, most of the accuracies are relatively flat or increase slightly until two to eight features remain. Therefore, for a particular classifier, most of the features are either irrelevant or redundant. Thus, if crowd reaction classification rates are to be improved significantly, other model classes or features should be used.

Feature selection methods that produce more accurate models with fewer features are assumed to have chosen better feature subsets. Consensus in selected features for well-performing models also indicates that the selected features are good predictors. This is explored more in Sec. IV.

2. Optimal time_steps_before

After feature selection, the hyperparameter *time_steps_before* was tuned again for the model, using the reduced feature set (see Fig. 7 and Sec. II D). Plots were made using the feature sets found by the Gini–Pearson method with the top three, four, four, and two features for *applause*, *chant*, *cheer*, and *distraction noise*, respectively, because the classification accuracies began to decrease more rapidly after these values.

The accuracies of *cheer* and *distraction noise* level out, while the accuracy of *chant* decreases after four previous half-second time steps are used. The models detecting *applause* continue to improve as they use more previous half-seconds. For the *chant* classifier with four features, the effect of including the feature temporal history of several time steps increased the accuracy more than 12%. Note that the optimal numbers of previous half-seconds found before feature selection were 12, 2, 3, and 4 for *applause, chant, cheer*, and *distraction noise*, respectively. The autocorrelation function for each feature up to 50 previous half-seconds was inspected in an attempt to explain this, but no correlations to optimal feature temporal history lengths were found.

To better understand the importance of feature temporal histories, the hyperparameter *time_steps_before* was also tuned with logistic regression models trained on all 27 features. The optimal numbers of time steps for each classifier were close to those found for the random forest, as seen in Table VII. The exception was *cheer*, which required a much longer temporal history than the random forest. LASSO regression was run on these expanded feature sets to gain insights into the role of each feature's temporal history.

TABLE V. The top five features for identifying crowd noise in order as ranked by each of the feature selection methods, given in order of most important to least important.

Gini	Gini–Pearson	Gini–Spearman	MDA	LASSO
PC 2	PC 2	PC 2	PC 2	Blue
$Leq_{0.5\ s}$	$Leq_{0.5\ s}$	Leq0.5 s	Leq0.5 s	PC 2
Blue	Cyan	Spectral decrease	Spectral entropy	Brown
Brown	Brown	Zero-crossing rate	Red	$Leq_{0.5\ s}$
Zero-crossing rate	Zero-crossing rate	Harmonic ratio	Spectral flatness	Zero-crossing rate

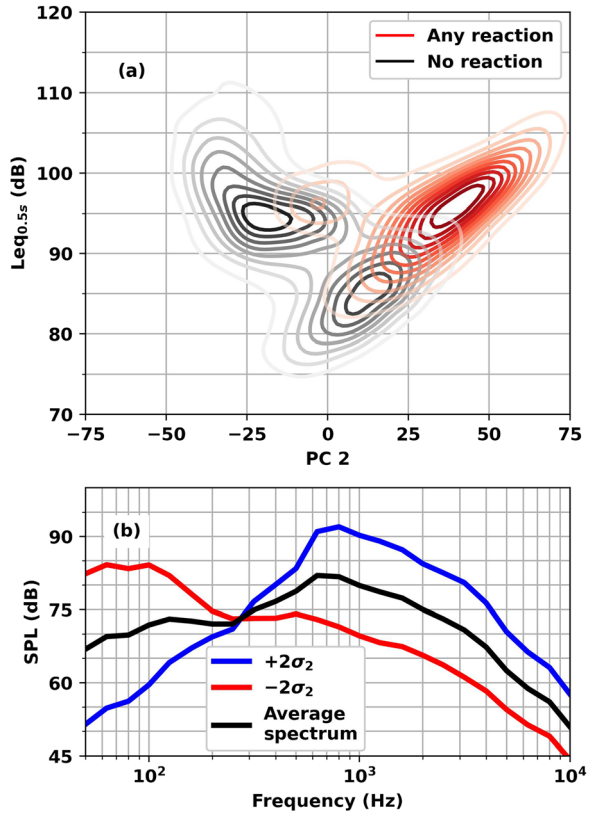

FIG. 5. Panel (a) shows a contour plot of the distribution of crowd reactions, which can be distinguished from other noises using the $Leq_{0.5\ s}$ as well as a principal spectral shape (PC 2) (Ref. 18) used to distinguish between sounds with high- and low-frequency content. For reference, the effect of PC 2 on the average 1/3-octave band spectra for indoor sports is shown in panel (b). The $\pm 2\sigma_2$ lines show the spectrum with PC 2 coefficients chosen 2 standard deviations away from where PC 2 is zero.

For *applause*, LASSO dropped intermediate half-seconds of the blue (high crowd involvement) and black (individual noise) cluster membership probability features[18] but ranked their oldest time steps highly. This reflects the physical context—at a basketball game, cheering often decays into applause instead of ending abruptly.

For *chant*, LASSO ranked the temporal history of the orange cluster membership probability feature (moderate crowd noise) as more useful than any other feature at any half-second time step. This is consistent with the physical context as well. Chanting is distinguished by a repeating pattern in a period of a few seconds.

TABLE VI. Top features and optimized parameter values for each classifier, as chosen by the Gini–Pearson feature selection. All models had $\log_2$ *max_features* and 3000 trees.

Classifier	Top features	Optimized parameters
Any reaction	PC 2 $Leq_{0.5s}$	*time_steps_before*: 2 *max_samples:* 0.68
Applause	Spectral flatness Blue Black	*time_steps_before*: 12 *max_samples:* 0.66
Chant	Spectral flatness Spectral decrease Black Yellow	*time_steps_before*: 2 *max_samples:* 0.66
Cheer	$Leq_{0.5\ s}$ Spectral centroid Brown Cyan	*time_steps_before*: 3 *max_samples:* 0.74
Distraction noise	Spectral entropy Blue	*time_steps_before*: 4 *max_samples:* 0.7

IV. DISCUSSION

In this section, physical interpretations are given for the results to draw out the connection between explainable machine learning and the subsets of features that were selected. The characteristics of the feature subsets are also analyzed for how they affect the number of previous half-second data points that optimize each model's performance.

A. Evaluation of features and *time_steps_before* selected for classifying reactions

1. Applause

Applause is generally characterized by collective clapping throughout the crowd. It is not rhythmically coordinated, but is sometimes accompanied by cheering. The most important features for classifying *applause* are shown in Table VIII. The top feature for *applause* was consistently spectral flatness, which is often reported in decibels.[33] Once converted to decibels, it takes on values from negative infinity (a pure tone) to zero (white noise) and represents the flatness of the linear spectrum.

Individual human clapping behavior follows several different modes, and each mode has a broad acoustic spectrum with different peaks[57] so that the overall effect of *applause*—the combined nonrhythmic clapping of many people—is broadband noise. This causes the distribution of spectral flatness for *applause* to have a higher mean than the other crowd reactions, as seen in Fig. 8(a). The second peak in the *chant* distribution to the right of *applause* can be explained by intermittent clapping during chants, which was verified by listening to sections of the game with chanting and comparing with plots of the spectral flatness. Some of the overlap between *applause* and *cheer* can be explained by instances where there is both *cheer* and *applause*. This overlap is also shown in Fig. 8(a).

The other feature that appeared in the top two features in two of the selection methods for distinguishing *applause* was the blue cluster membership probability.[18] The blue cluster is highly correlated with PC 2 [Fig. 2(a)], another highly ranked feature. The blue cluster (and to a lesser degree, PC 2) is also highly correlated with *distraction noise*, as seen in Fig. 2(b). Density plots of spectral flatness against the blue cluster 13 steps before, 13 steps after, and concurrently [Figs. 8(b)–8(d)] show that the temporal relationship between the two features is such that when identifying *applause*, it is more useful to know the blue cluster membership probability concurrent with or before spectral flatness, rather than knowing the spectral flatness before the blue cluster membership probability.

https://doi.org/10.1121/10.0039709

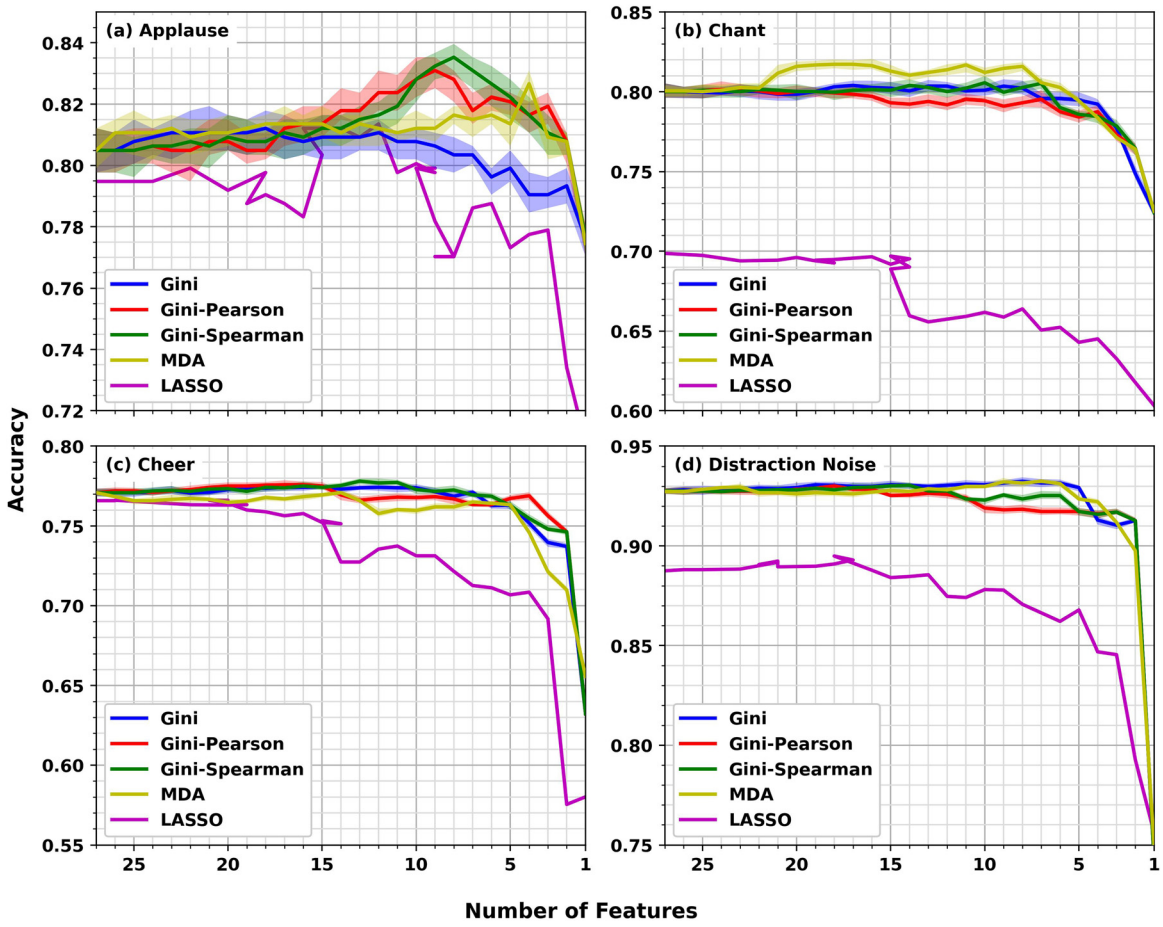

FIG. 6. The accuracy of binary classification models for specific crowd reactions as a function of the number of features. Transparent bands show the minimum and maximum accuracies after running the process with 20–50 different random seeds. Bands in panels (b)–(d) are difficult to see because they tightly follow the mean accuracy curves.

Applause often follows game events that are positive, but not as significant, such as a turnover by the opposing team. It also occurs as cheering dies down. Both *cheer* and *distraction noise* (made by fans of the home team before an opposing team turnover) are correlated with blue, so it is possible that the classifiers detect *applause* not only by the sound of *applause* itself, but also by the crowd reactions that it commonly follows. Hence, spectral flatness and the blue cluster membership probability are acoustic features that are not only useful for machine learning, but are also connected to the physical processes of *applause* and crowd sounds that often precede *applause*. This also helps explain why the classifier requires a higher *time_steps_before* to predict applause than to predict the other three crowd reactions.

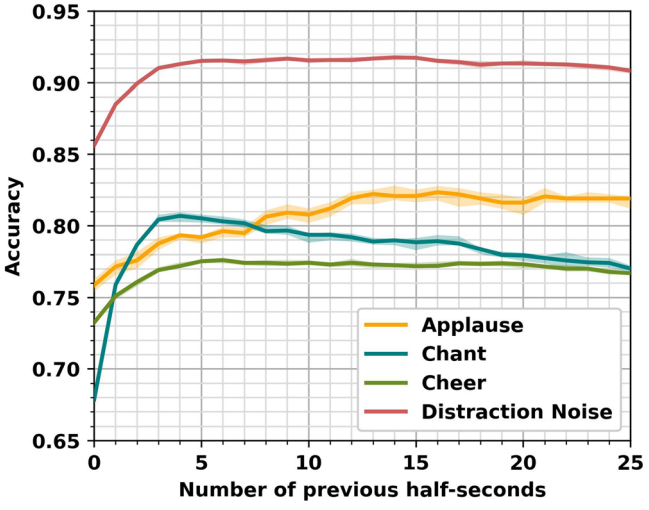

FIG. 7. The accuracy of crowd reaction classification models as a function of *time_steps_before* using the reduced feature set found by the Gini–Pearson feature selection method.

2. Chant

Chants are semi-organized rhythmic vocalizations that are carried out by crowds in unison, sometimes interspersed

TABLE VII. The optimal number of previous time steps for logistic regression vs random forests (before feature selection).

Reaction	Random forest	Logistic regression
Any reaction	2	6
Applause	12	16
Chant	2	4
Cheer	3	25
Distraction noise	4	4

TABLE VIII. The top five features for distinguishing between *applause* and other crowd reactions as ranked by each of the feature selection methods, given in order of most important to least important.

Gini	Gini–Pearson	Gini–Spearman	MDA	LASSO
Spectral flatness	Spectral flatness	Spectral flatness	Spectral flatness	Orange
Spectral entropy	Blue	PC 2	Blue	Spectral entropy
Spectral roll-off point	Black	Zero-crossing rate	Zero-crossing rate	Spectral flatness
Spectral centroid	Spectral decrease	PC 1	Green	Harmonic ratio
PC 3	Zero-crossing rate	Spectral skewness	Spectral roll-off point	PC 2

with moments of rhythmic clapping. The top feature for detecting *chant* was consistently spectral flatness, as seen in Table IX.

As suggested in Sec. IV A 1, spectral flatness increases during clapping, and chants often include intermittent clapping. Intermittent clapping can be seen by plotting spectral flatness against itself 2 half-seconds before, as seen in Fig. 9. The peaks around (−22, −12) and (−12, −22) indicate that over two time steps, the spectral flatness for *chant* is likely to shift between two values, while it is not likely to shift for *other reactions*. This was verified by comparing recordings of chants to plots of the spectral flatness. Similar distributions can be seen by looking at delays of 1, 3, and 4 half-seconds. This result is consistent with what was observed in Sec. III B 2, which showed that classifiers detecting *chant* improved by 12% when including several previous half-seconds. The random forest is possibly looking for a place where the spectral flatness switches between these two values in the recent feature temporal history. This also might explain why the LASSO logistic regression model performed much worse than the random forest models, since the correlation for a feature that switches between

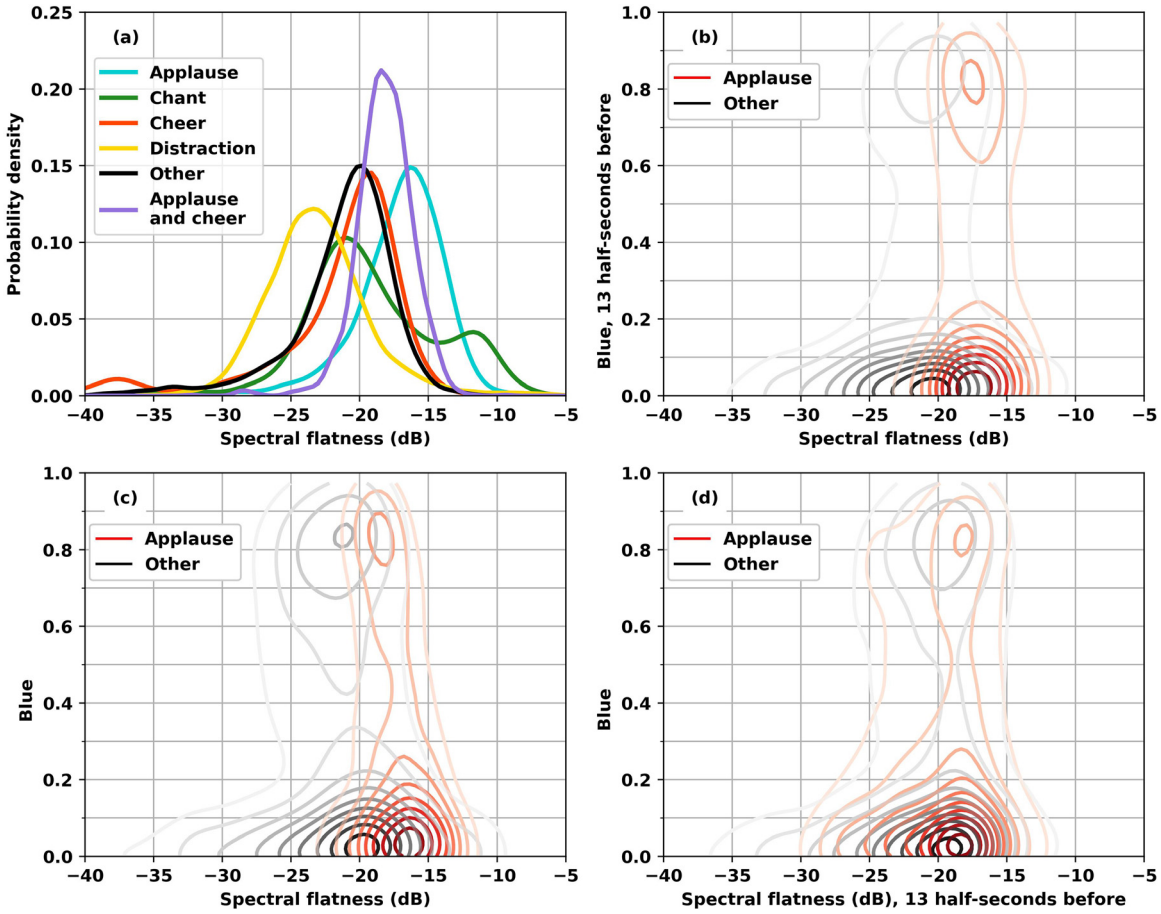

FIG. 8. In panel (a), the distribution of spectral flatness for several crowd reactions is shown using a Gaussian kernel density estimate. *Applause* is often accompanied by *cheer*, so the distributions for *only applause* and *only cheer* are shown, in addition to a distribution of half-seconds containing both *cheer* and *applause*. In panels (b)–(d), the distribution of spectral flatness and blue cluster membership probability is shown for *applause* and *other reactions*. A time delay for the feature is introduced in panels (b) and (d) for blue and spectral flatness, respectively, to show how the features have a temporal correlation when blue is delayed 13 half-seconds or when neither feature is delayed, but not when spectral flatness is delayed 13 half-seconds.

https://doi.org/10.1121/10.0039709

TABLE IX. The top five features for distinguishing between *chant* and other crowd reactions as ranked by each of the feature selection methods, given in order of most important to least important.

Gini	Gini–Pearson	Gini–Spearman	MDA	LASSO
Spectral flatness	Spectral flatness	Spectral flatness	Spectral flatness	Orange
PC 3	Spectral decrease	Spectral decrease	Spectral centroid	PC 3
Spectral centroid	Black	PC 2	Pink	Spectral crest
Spectral crest	Yellow	$Leq_{0.5\ s}$	Spectral entropy	Yellow
Spectral decrease	Orange	PC 3	Harmonic ratio	Pink

two values in time becomes nonlinear. Once again, there is a strong connection between the acoustic feature of spectral flatness and the acoustic properties of *applause*.

Other than spectral flatness, *chant* has the strongest Pearson correlations with orange, spectral spread, spectral entropy, spectral crest, and PC 3. Interestingly, three of these five features appear in the top five features as ranked by LASSO, even though other models performed significantly better than LASSO. This further suggests that the temporal relationships between variables are more important than correlations between features and reactions.

Spectral decrease appears three times in the top five features and twice as the second-most-important feature across all models. Of all the features from the MATLAB acoustic feature extractor, it was the least correlated with other features. This might have in part been due to the feature's hyperbolic arcsine transformation, but the other feature that had a hyperbolic arcsine transformation (spectral skewness) was highly correlated with several other features. Spectral decrease measures the slope of the spectrum (like spectral slope), but emphasizes lower frequencies more. These lower frequencies may span the voice fundamental frequencies of individuals chanting in the crowd. During joint speech activities, such as *chant*, individuals tend to converge toward a shared fundamental frequency,[58,59] which may have been what increased this feature's importance.

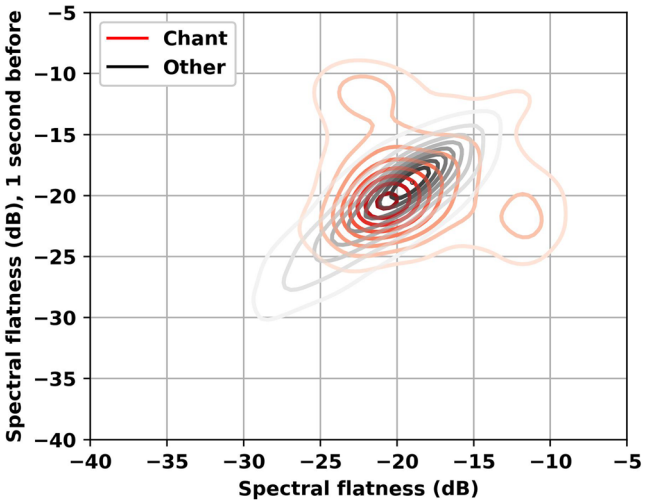

FIG. 9. The distribution of *chant* and *other reactions* across the feature spectral flatness and the spectral flatness 2 half-seconds before. The peaks near (−12, −22) and (−22, −12) show that one of the characteristics of *chant* is a switching between high and low spectral flatness on short time intervals. A similar shape can be seen by observing the distribution of *chant* over spectral flatness and spectral flatness from 3 or 4 half-second time steps before.

3. Cheer

Cheer is generally characterized by loud, positive vocalizations with little rhythmic structure and intermittent coordination between individuals in the crowd. The most important feature appeared to be the $Leq_{0.5\ s}$. The spectral centroid was also ranked highly by two high-performing feature selection methods (see Table X). Out of all the crowd reactions, *cheer* has the most acoustic energy, which explains why the $Leq_{0.5\ s}$ might be a useful feature for identifying *cheer*. It also has a spectral centroid that is shifted to higher frequencies because cheering is generally a happy reaction to a game event. Happy/positive emotions tend to be vocalized with higher pitches (higher fundamental frequencies), and the constriction of the vocal tract also heightens the frequencies in the first two formants compared to neutral or negative emotions.[60] The relationship between spectral centroid and the $Leq_{0.5\ s}$ for *cheer* appears in Fig. 10 and shows some of the simple, though incomplete, logic in feature selection behind the complicated random forest model used to identify *cheer*.

4. Distraction noise

Distraction noise is commonly characterized by a crowd vocalizing in unison, often unifying toward a single long tone with little rhythmic variation. It often occurs during a free-throw or during play when the other team has possession of the ball, when other sounds (like the band or PA system) are not present. The *distraction noise* classifier had the highest accuracy and the most agreement among the feature selection methods, with only MDA and LASSO not ranking the blue cluster membership probability as the second-most-important feature, as seen in Table XI. MDA ranked the blue cluster membership probability as the least important feature, and this could be attributed to the high variance in individual feature rankings in the early stages of MDA. The blue cluster was identified in Ref. 18 to be highly correlated with *distraction noise*, and this is further confirmed by Fig. 2. The average 1/3-octave half-second spectrum for the blue cluster in shown in Fig. 11(b). The spectrum has large factors of the first and second principal spectral shapes (PC 1 and PC 2),[18] which correspond to loud, crowd-dominated sounds.

TABLE X. The top five features for distinguishing between *cheer* and other crowd reactions as ranked by each of the feature selection methods, given in order of most important to least important.

Gini	Gini–Pearson	Gini–Spearman	MDA	LASSO
$Leq_{0.5\ s}$	$Leq_{0.5\ s}$	$Leq_{0.5\ s}$	Spectral flatness	$Leq_{0.5\ s}$
Spectral entropy	Spectral centroid	Spectral centroid	PC 3	Brown
Short time energy	Brown	Spectral decrease	Spectral flux	PC 2
Brown	Cyan	Spectral skewness	Pink	Harmonic ratio
Spectral centroid	Red	PC 3	Blue	PC 1

Spectral entropy was the other highly ranked feature for classifying *distraction noise*. It measures the entropy of the linear power spectrum by treating the spectrum as a distribution. *Distraction noise* is the most tonal of crowd reactions, and tonal sounds tend to have lower spectral entropy. The distribution of *distraction noise* and *other reactions* can be seen in Fig. 11, which shows a fairly clear distinction between the distribution of *distraction noise* and other crowd reactions. Hence, the feature selection process revealed that *distraction noise* can be identified by machine learning models using features that correspond to tonality with a specific spectral shape corresponding to loud sounds made by crowds.

B. Performance of feature selection methods on crowd noise data

The random forest-based feature selection methods used in this paper generally had good consensus on what the most important feature is, and reasonable candidates for the second-best feature could often be found by looking at the highest-ranked features from all the methods. As expected, a lot of the top features found by the Gini importance metric were highly correlated with each other. MDA had the highest variance in its accuracy across the feature selection process. This might have been mitigated by using more random seeds in the feature selection process, especially during the first few steps, where the process appeared to be more random. However, this would increase the cost of an already expensive computation. LASSO logistic regression often had lower accuracy than the other methods, which shows that the temporal aspect of crowd noise may be more important than finding features that have high linear correlations with crowd reactions. An alternative explanation is random forests are better at learning nonlinear relationships among features and labels.

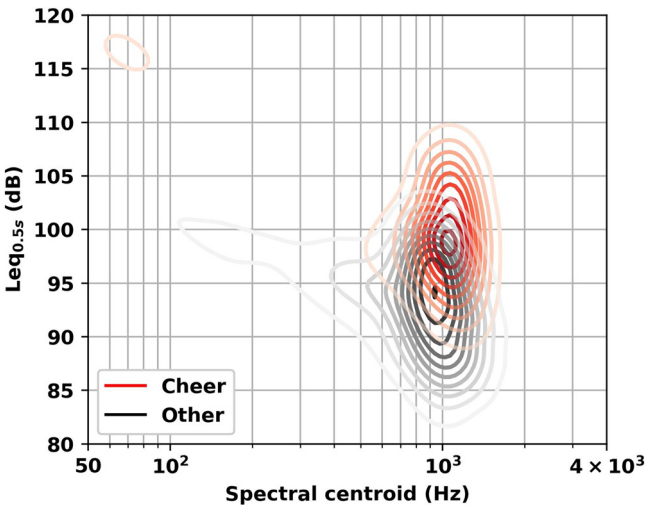

FIG. 10. The distribution of $Leq_{0.5\ s}$ and spectral centroid for *cheer* and other crowd reactions. *Cheer* has a higher $Leq_{0.5\ s}$ and spectral centroid than *other reactions*.

Models found using the Gini feature importance metric with a correlation penalty had best overall accuracy, with the Pearson correlation penalty performing slightly better than the Spearman correlation penalty. The correlation penalty method consistently found good candidates for the two most relevant features for classifying each crowd reaction, as measured by the model accuracy. This might have been because it was less dependent on the random seed and encouraged a weakly correlated feature subset to be chosen. Although models produced using Gini feature importance with a correlation penalty had slightly higher accuracy on average than models produced with other importance metrics, more research would need to be done to determine whether this generalizes to other machine learning datasets and problems.

C. Limitations of the study

Reverberation times in college basketball stadiums vary as much as 3 s.[61] Because features in this study were calculated every half-second, sound pressure features for the same crowd reactions may look significantly different at venues with different reverberation times. This may impact the models' ability to predict crowd reactions. If these models are to be used in situations other than the basketball stadium considered in this study, such as an outdoor football stadium, the impact of reverberation times must be understood.

Because data were labeled by multiple human labelers, the reported accuracy of each model is limited by the consistency between labelers. One particularly difficult task was distinguishing between crowd noise and noise from individuals to decide when a crowd reaction begins and ends. Though labelers met regularly to coordinate on questions like these, human labeling inevitably introduced ambiguity into labeled data for each reaction. While labelers did meet periodically to ensure similar labeling decisions were being

https://doi.org/10.1121/10.0039709

TABLE XI. The top five features for distinguishing between *distraction noise* and other crowd reactions as ranked by each of the feature selection methods, given in order of most important to least important.

Gini	Gini–Pearson	Gini–Spearman	MDA	LASSO
Spectral entropy	Spectral entropy	Spectral entropy	Black	Blue
Blue	Blue	Blue	Spectral entropy	Harmonic ratio
Harmonic ratio	Black	Black	$Leq_{0.5\,s}$	PC 1
Black	Orange	Spectral decrease	PC 3	Pink
Spectral flatness	Spectral decrease	Orange	Spectral crest	Spectral crest

made, this consistency between labelers was not systematically checked.

Because *cheer* often turns into *applause* as the crowd calms down, labels for *cheer* and *applause* frequently overlapped. As a result, the features selected for identifying *cheer* may be influenced by the overlapping *applause*, and vice versa. Data with overlapping labels represent 17.5% of data labeled *cheer* and 31.7% of data labeled *applause*. Despite this overlap, the best feature selection method, Gini–Pearson, did not choose any of the same top five features for *cheer* and *applause*.

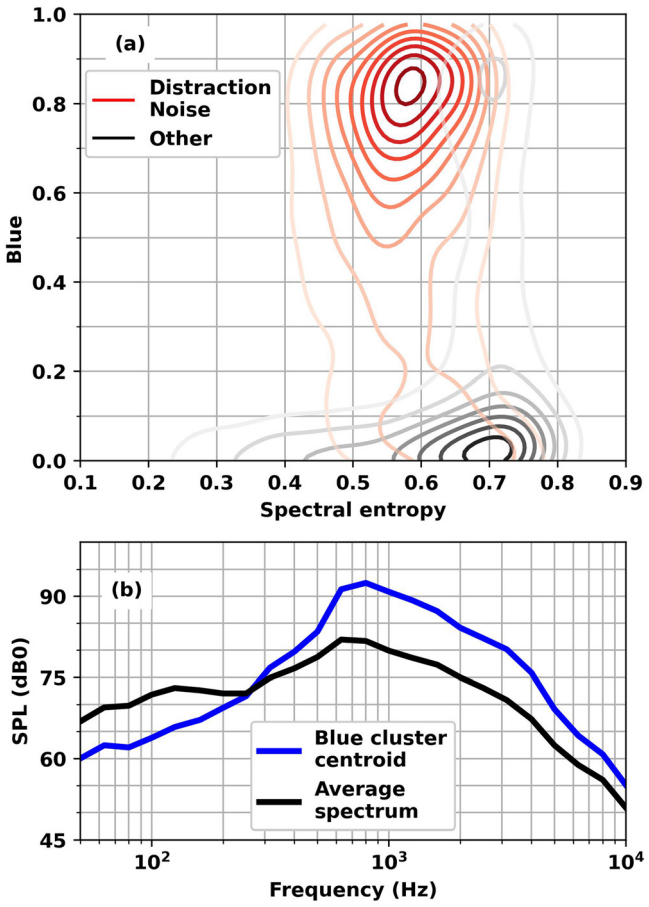

FIG. 11. *Distraction noise* is the most easily identified of crowd reactions. Panel (a) shows the distribution of *distraction noise* and *other reactions* across the features blue cluster membership probabilities and spectral entropy. Panel (b) shows the spectral shape of the blue cluster centroid, which is highly correlated with *distraction noise* and contains much more high-frequency content than the average 1/3-octave spectrum for indoor sporting events, which is shown for reference.

V. CONCLUSION

This work identified several acoustic features that are relevant for classifying crowd reactions at collegiate men's basketball games. Specifically, the $Leq_{0.5\,s}$ and a principal spectral shape found in Ref. 18 relating to the amount of high-frequency content in half-second 1/3-octave band spectra (abbreviated as PC 2 in this paper) are consistently ranked as important features for identifying any crowd reaction, while features relating to tonality, specific spectral shapes, and sound level are useful for distinguishing between crowd reactions. These features often have explainable connections to physical phenomena such as clapping, loud crowds, and the tonality of certain sounds.

The importance of feature temporal histories in crowd reaction classification was also quantified. For example, *chant* classification accuracy improved by 12% when the current and previous 4 half-second steps were included in the feature set to identify intermittent clapping, among other temporal patterns. *Applause* can be identified by transitions from spectral shapes relating to *distraction noise* or *cheer* to high spectral flatness, which relates to broadband clapping noise.

Of all the reduced feature models generated, the two most promising models were derived from the Gini feature importance metrics with correlation penalties. These models are consistently better classifiers on this dataset than the other models with similarly small feature vectors.

The acoustic features identified here can be used as a starting point for future research. These features may be used with the labeled crowd responses with different machine learning algorithms in an effort to achieve higher classification accuracies. Additionally, although this study identifies that several half-seconds of feature temporal history improves model performance for identifying most crowd reactions, the exact relationship between various features and crowd reactions and the evolution of those relationships through time still remain unclear. Further analysis of these features (e.g., by allowing variations in the number of previous half-seconds of feature temporal history on a per-feature basis) may yield useful insights into transitions between different forms of crowd behavior. It may also be that different acoustic features are important for different sporting events, depending on the physical properties of the venue and the size and demographics of the crowd, as well as behavioral norms around acoustic crowd reactions for different sports. The features identified here can guide the

selection of a reasonable starting feature set or identify features that may improve models across different sports. Improvements in the ability to classify crowd reactions in a context-specific setting—such as sporting events, which come with certain expected norms of acoustic behavior from crowds—may lay the groundwork for identifying important acoustic features and building models to classify these features for different types of audience or within less-scripted crowd behaviors. Such models can potentially inform venues on the sentiment/mood of a crowd, quantify the value of individual players for name, image, and likeness contracts, or aid decision making for emergency response.

In conclusion, this work identified acoustic features that are particularly relevant for classifying crowd behavior at basketball games. Of the feature selection methods used in this study, those using Gini feature importance metrics with correlation penalties yielded the most promising results. This paper also showed that incorporating acoustic information about previous crowd responses can help classify current crowd reactions. This work provides a foundation for the classification of reactions within crowd noise at other events.

ACKNOWLEDGMENTS

The authors thank Christian Anderson for helpful conversations. This work was supported by the Brigham Young University College of Computational, Mathematical, and Physical Sciences through the College High-Impact Research Program.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1]C. Horbel, B. Popp, H. Woratschek, and B. Wilson, "How context shapes value co-creation: Spectator experience of sport events," Serv. Ind. J. **36** (11–12), 510–531 (2016).
[2]N. Popp, C. McEvoy, and N. Watanabe, "Do college athletics marketers convert social media growth into ticket sales?," Int. J. Sports Mark. Sponsorship **18**(2), 212–227 (2017).
[3]J. P. Doyle, Y. Su, and T. Kunkel, "Athlete branding via social media: Examining the factors influencing consumer engagement on Instagram," Eur. Sport Manag. Q. **22**(4), 506–526 (2022).
[4]NCAA, "Board of governors starts process to enhance name, image and likeness opportunities," https://www.ncaa.org/news/2019/10/29/board-of-governors-starts-process-to-enhance-name-image-and-likeness-opportunities.aspx (2019) (Last viewed July 5, 2024).
[5]T. Kunkel, B. J. Baker, T. A. Baker, and J. P. Doyle, "There is no nil in NIL: Examining the social media value of student-athletes' names, images, and likeness," Sport Manag. Rev. **24**(5), 839–861 (2021).
[6]A. R. Cocco and A. M. Moorman, "Untapped potential: An examination of name, image, and likeness earnings estimates for community college athletes," J. Issues Intercollegiate Sports **15**(1), 256–271 (2022).
[7]R. Grohs, V. E. Wieser, and M. Pristach, "Value cocreation at sport events," Eur. Sport Manag. Q. **20**(1), 69–87 (2020).
[8]G. Baranowski-Pinto, V. L. S. Profeta, M. Newson, H. Whitehouse, and D. Xygalatas, "Being in a crowd bonds people via physiological synchrony," Sci. Rep. **12**(1), 613 (2022).
[9]S. Proksch, M. Reeves, K. Gee, M. Transtrum, C. Kello, and R. Balasubramaniam, "Recurrence quantification analysis of crowd sound dynamics," Cogn. Sci. **47**(10), e13363 (2023).
[10]V. Franzoni, G. Biondi, and A. Milani, "Emotional sounds of crowds: Spectrogram-based analysis using deep learning," Multimed. Tools Appl. **79**(47–48), 36063–36075 (2020).
[11]N. Singh, N. Roy, and A. Gangopadhyay, "Analyzing the sentiment of crowd for improving the emergency response services," in *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, Taormina, Italy (IEEE, New York, 2018), pp. 1–8.
[12]S. Reicher, C. Stott, P. Cronin, and O. Adang, "An integrated approach to crowd psychology and public order policing," Policing **27**, 558–572 (2004).
[13]M. Hayne, R. Rumble, and D. Mee, "Prediction of crowd noise," in *ACOUSTICS 2006: First Australasian Acoustical Societies' Conference*, Christchurch, New Zealand (Australian Acoustical Society, Magill North, SA, Australia, 2006), pp. 235–240.
[14]M. Hayne, J. Taylor, R. Rumble, and D. Mee, "Prediction of noise from small to medium sized crowds," in *ACOUSTICS 2011: Australian Acoustical Society Conference*, Gold Coast, Australia (Australian Acoustical Society, Magill North, SA, Australia, 2011), pp. 686–692.
[15]A. Barnard, S. Porter, J. Bostron, R. terMeulen, and S. Hambric, "Evaluation of crowd noise levels during college football games," Noise Control Eng. J. **59**(6), 667–680 (2011).
[16]K. Hummel, E. Ryherd, X. Cheng, and B. Lowndes, "Relating clustered noise data to hospital patient satisfaction," J. Acoust. Soc. Am. **154**(2), 1239–1247 (2023).
[17]M. Sano, H. Sumiyoshi, M. Shibata, and N. Yagi, "Generating metadata from acoustic and speech data in live broadcasting," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, PA (IEEE, New York, 2005), Vol. 2, pp. ii/1145–ii/1148.
[18]M. C. Cutler, M. R. Cook, M. K. Transtrum, and K. L. Gee, "Data-driven decomposition of crowd noise from indoor sporting events," J. Acoust. Soc. Am. **155**(2), 962–970 (2024).
[19]B. A. Butler, K. Pedersen, M. R. Cook, S. G. Wadsworth, E. Todd, D. Stark, K. L. Gee, M. K. Transtrum, and S. Warnick, "Classifying crowd behavior at collegiate basketball games using acoustic data," Proc. Mtgs. Acoust. **35**(1), 055006 (2018).
[20]E. Todd, M. R. Cook, K. Pedersen, D. S. Woolworth, B. A. Butler, X. Zhao, C. Liu, K. L. Gee, M. K. Transtrum, and S. Warnick, "Automatic detection of instances of focused crowd involvement at recreational events," Proc. Mtgs. Acoust. **39**(1), 040003 (2020).
[21]R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," J. Appl. Sci. Technol. Trends **1**(1), 56–70 (2020).
[22]M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," in *Computational Intelligence and Bioinspired Systems: IWANN 2005*, edited by J. Cabestany, A. Prieto, and F. Sandoval (Springer, Berlin, Germany, 2005), pp. 758–770.
[23]J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," ACM Comput. Surv. **50**(6), 1–45 (2017).
[24]L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML'03: Proceedings of the 20th International Conference on Machine Learning*, Washington, DC (Association for Computing Machinery, New York, 2003), pp. 856–863.
[25]J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, "Designing a feature selection method based on explainable artificial intelligence," Electron. Markets **32**(4), 2159–2184 (2022).
[26]BYU Athletics, "Men's basketball archived season stats—BYU Athletics—Official Athletics website—BYU Cougars," byucougars.com/mens-basketball-archived-season-stats (Last viewed October 18, 2025).
[27]MathWorks, "MATLAB version 9.12.0 (R2022a) [computer program]," www.mathworks.com (2024) (Last viewed June 28, 2024).

[28]MathWorks, "Audio Toolbox version 3.2 (R2022a) [computer program]," www.mathworks.com (2024) (Last viewed June 28, 2024).
[29]MathWorks, "audioFeatureExtractor [computer program]," https://www.mathworks.com/help/audio/ref/audiofeatureextractor.html (2019) (Last viewed June 28, 2024).
[30]MathWorks, "pitch [computer program]," https://www.mathworks.com/help/audio/ref/pitch.html (2018) (Last viewed June 28, 2024).
[31]G. Peeters and X. Rodet, *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project* (IRCAM, Paris, France, 2004).
[32]H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada (IEEE, New York, 2004), Vol. 1, pp. I-193.
[33]J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," IEEE J. Sel. Areas Commun. **6**(2), 314–323 (1988).
[34]E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany (IEEE, New York, 1997), Vol. 2, pp. 1331–1334.
[35]A. Lerch, *An Introduction to Audio Content Analysis Applications in Signal Processing and Music Informatics* (IEEE Press, Piscataway, NJ, 2012).
[36]H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval* (John Wiley & Sons, Hoboken, NJ, 2006).
[37]M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," J. Acoust. Soc. Am. **146**(5), 3590–3628 (2019).
[38]J. C. De Winter, S. D. Gosling, and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," Psychol. Methods **21**(3), 273–290 (2016).
[39]J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data," Quaestiones Geographicae **30**(2), 87–93 (2011).
[40]L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).
[41]L. Breiman, *Manual on Setting up, Using, and Understanding Random Forests v3.1* (University of California, Berkley, 2002), Vol. 1, p. 58.
[42]K. Pedersen, M. K. Transtrum, K. L. Gee, S. V. Lympany, M. M. James, and A. R. Salton, "Feature selection for a continental-scale geospatial model of environmental sound levels," J. Acoust. Soc. Am. **154**(2), 1168–1178 (2023).
[43]A. M. Musolf, E. R. Holzinger, J. D. Malley, and J. E. Bailey-Wilson, "What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics," Hum. Genet. **141**(9), 1515–1528 (2022).
[44]H. Wang, F. Yang, and Z. Luo, "An experimental study of the intrinsic stability of random forest variable importance measures," BMC Bioinformatics **17**, 60 (2016).
[45]G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," Stat. Comput. **31**, 82 (2021).
[46]X. Li, Y. Wang, S. Basu, K. Kumbier, and B. Yu, "A debiased MDI feature importance measure for random forests," Adv. Neural Inf. Process. Syst. **32**, 1–22 (2019), arXiv:1906.10845v2.
[47]K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," Comput. Stat. Data Anal. **52**(4), 2249–2260 (2008).
[48]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[49]P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," Wiley Interdiscip. Rev. **9**(3), e1301 (2019).
[50]P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," J. Mach. Learn. Res. **18**(181), 1–18 (2018).
[51]J. Humpherys and T. J. Jarvis, *Foundations of Applied Mathematics, Volume 2: Algorithms, Approximation, Optimization* (SIAM, Philadelphia, PA, 2020), p. 142.
[52]A. Liaw and M. Wiener, "Classification and regression by randomForest," R News **2**(3), 18–22 (2002).
[53]H. Deng, G. Runger, and E. Tuv, "Bias of importance measures for multivalued attributes and solutions," in *Artificial Neural Networks and Machine Learning* (Springer, New York, 2011), Vol. 2, pp. 293–300.
[54]C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics **8**, 25 (2007).
[55]T. Hastie, G. James, R. Tibshirani, and D. Witten, *An Introduction to Statistical Learning* (Springer, New York, 2021), pp. 241–242.
[56]B. Hoyle, M. M. Rau, R. Zitlau, S. Seitz, and J. Weller, "Feature importance for machine learning redshifts applied to SDSS galaxies," Mon. Not. R. Astron. Soc. **449**(2), 1275–1283 (2015).
[57]B. H. Repp, "The sound of two hands clapping: An exploratory study," J. Acoust. Soc. Am. **81**(4), 1100–1109 (1987).
[58]V. Aubanel and N. Nguyen, "Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task," PLoS One **15**(5), e0232209 (2020).
[59]A. R. Bradshaw and C. McGettigan, "Convergence in voice fundamental frequency during synchronous speech," PLoS One **16**(10), e0258747 (2021).
[60]R. G. Kamiloğlu, A. H. Fischer, and D. A. Sauter, "Good vibrations: A review of vocal expressions of positive emotions," Psychon. Bull. Rev. **27**, 237–265 (2020).
[61]M. Shepherd, S. A. Hambric, N. D. Evans, D. J. Domme, A. W. Christian, B. P. Cranage, K. Poulain, A. J. Orr, A. R. Barnard, and M. D. Gardner, "Rating of the loudest college basketball arenas for ESPN magazine," Proc. Mtgs. Acoust. **12**(1), 015004 (2011).