IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 10, 2021, accepted July 7, 2021, date of publication July 13, 2021, date of current version July 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096776

Person Identification From Audio Aesthetic

BRANDON SIEU, (Member, IEEE), AND MARINA L. GAVRILOVA, (Senior Member, IEEE)
Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada

Corresponding author: Brandon Sieu (brandon.sieu@ucalgary.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC), and in part by the Innovation for Defence Excellence and Security (IDEaS).

ABSTRACT Behavioral biometrics survey actions rather than the physical traits of the person. Within this categorization, social behavioral biometrics utilizes an individual's communications for biometric analysis. The investigation of the uniqueness of human preferences and their implications to other aspects of an individual, such as personality or gender, is both a psychological and a biometric problem. An emerging approach is the usage of an individual's aesthetic preferences for the purpose of person identification. Recent research into the identification from visual aesthetics has found that these preferences hold significant discriminatory value. However, aesthetic identification has only been conducted through a visual medium via a set of liked images. The contribution of this work is the development of the first audio aesthetic preference system for person identification. The proposed system extracts descriptive intra-song and inter-song features from a set of songs favored by users and utilizes an ensemble of classifiers for prediction. The final decision is optimized by a genetic algorithm. Experimental results demonstrate that the developed audio aesthetic system achieves 95% user recognition accuracy on both proprietary and public audio datasets.

INDEX TERMS Audio aesthetics, behavioral biometrics, biometric security, human–machine interactions, pattern recognition.

I. INTRODUCTION

Biometric analysis investigates the physical aspects of a person. Well researched domains include fingerprint identification, iris analysis, and facial recognition. Behavioral biometrics is a subset of biometrics that inspects an individual's actions rather than their physical traits. This form of biometrics can be used to analyze a person covertly and remotely. Within the area of behavioral biometrics, social behavioral biometrics studies the interactions, attitudes, and communications of a person [1]. Social behavior is especially prominent in the modern online spheres, where social networks and platforms allow for widespread public communication. With an individual's social behavioral features, inference or identification systems can be implemented that do not require any physical contact with the user.

A recent direction of social behavioral biometric research is exploring the use of aesthetic preference as features for person identification. Aesthetic preference can be described as an individual's likeness or fondness of subject material. The term has been traditionally used in the context of art, pertaining to one's judgment or taste in beauty [2]. However, the concept of aesthetic preference has been broadened to an immediate pleasurable experience toward an object [3]. Preference information is widely available on online platforms that utilize a subscription or endorsement system.

Research into the identification of users through visual aesthetic preference is a new domain. Given a set of favorite images, systems have been developed that can extract discriminatory features from this set to accurately identify users [4]. From these works, it was determined that there are unique qualities to a person's visual aesthetics that can be used for the problems of person identification and gender identification [5], [6]. Although the area is still emerging, the use of human aesthetic information as features shows high potential. The advantage of aesthetic data is that it is both retrievable through online systems and does not require active participation from the user. Such systems can be extended to understanding consumer behavior, tailoring personal user experiences, and gaining insights into the unique properties of human aesthetics.

Over the past two years, there has been an emergence in research on the analysis of music through advanced machine learning. This includes recent IEEE access publications studying the use of deep neural networks in the surveillance of roads from anomalous sounds [7], the generation of high-fidelity audio samples through adversarial autoencoders [8], and new approaches of music genre

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

classification [9], [10]. Although the audio analysis domain is well researched, this paper proposes a new topic of audio aesthetics for identification. The discriminatory value of an individual's musical aesthetics has not been explored, while recent aesthetic research has achieved accurate results using only visual aesthetics [4].

The fundamental premise is that a person's audio preference is unique and holds discriminatory value. Many factors can influence an individual's musical taste, ranging from cultural background to personal experience [11]. Furthermore, research exists that links musical preference to more physiological aspects of an individual such as cognitive function and emotion [12]. Exploring the correlations between audio preference and social behavior, such as personality, is an emerging research direction [13].

The following research questions are asked:

- Do audio aesthetic features hold discriminatory value and can audio aesthetics be used for user identification?
- Can a system be developed to accurately identify users only using information from songs in a set related to the same user?
- How can the relationship between the songs in a set be modeled for user prediction?

In this paper, we propose and develop the first, to the best of our knowledge, person identification system using audio aesthetics. Audio data varies greatly from visual data, and therefore, different methodologies for observing and processing the signal's properties along the time domain must be used. The novel research direction of using a person's audio aesthetic preference for identification is a natural progression of prior research. For this purpose, two new datasets were collected consisting of songs liked by the users.

The paper makes the following contributions to address the research questions:

- A system has been developed to address the emerging problem of person identification from audio aesthetics. To the best of our knowledge, this is the first work that established the discriminatory value of audio aesthetic features.
- A new methodology of combining classical audio features (intra-song features) with cross-similarity features (inter-song features) is proposed. Utilizing the inherent knowledge that the samples within a given set belong to the same user, features between each pair of songs are extracted to quantify the user's dynamic song preference range.
- The relationship between songs in a set is modeled through the segmentation and concatenation of the songs in the set. Each signal is separated using Harmonic Percussive Source Separation (HPSS), and further segmented with a time window for feature extraction. The data is then concatenated to form a 1D representation of a song set for user prediction.

Two audio aesthetic datasets resembling the structure of the original visual aesthetic dataset [5] are constructed for comparison. These two datasets are composed of songs sampled from the Free Music Archive (FMA) and the Million Playlist Dataset (MPD), respectively. With the information extracted from these song sets, the system can identify users with 95.74% rank 1 accuracy. Moreover, results on the proprietary dataset demonstrate that combining a visual and an audio system using score-level fusion leads to overall recognition accuracy of 99.41%. The developed system can find potential applications in recommendation systems, multi-factor authentication, and consumer behavioral analysis.

II. RELATED WORKS

Aesthetic identification is an emerging research domain. The first proof of concept research was carried out by Lovato *et al.* [5] in 2012. In this work, a dataset consisting of 200 users each with 200 liked images was collected from the image hosting website Flickr. Various features were extracted from the images, including color, edges, textures, regions, objects, faces, and scene features. Sparse regression using Least Absolute Shrinkage and Selection Operator (LASSO) regularization was used to determine the feature weight. The recognition accuracy of the method was inferior to later works, but this proof of concept concluded that there was at least some discriminatory value in the images picked by each user.

In 2014, Segalin *et al.* [14] proposed a method utilizing a multi-resolution counting grid to train an ensemble of Support Vector Machine (SVM) classifiers in a one-versus-all approach. Bags of Features were used with 111 features and converted into intensity maps for each user. These maps are then used as inputs into the multi-resolution counting grid. A rank 1 accuracy of 73% was found using 100 training images. The study showed that machine learning approaches and additional features could improve accuracy significantly.

By introducing new feature categories to the images and with more sophisticated feature engineering, Azam and Gavrilova [15] obtained a rank 1 accuracy of 84% on the same dataset. The features were categorized into four distinctions: local/global perceptual features, HOG features, and content features with 861 total features. After applying Principal Component Analysis (PCA), LASSO regression was used for identification.

Following this trend of feature engineering, Brandon and Gavrilova [16] used Gene Expression Programming (GEP) to construct 150 complex features from an original 924 features. The approach utilized evolutionary programming with the objective of creating dense, discriminatory combinations within the feature set. An initial population is produced which is modified through natural selection with classification accuracy as the objective function. These final dense features decreased the memory requirements for the system and allowed for more accurate classification. A rank 1 accuracy of 94.1% was achieved.

Deep learning-based systems have been developed that range from the detection of contextual situations to the efficient recognition of faces [17], [18]. Although features

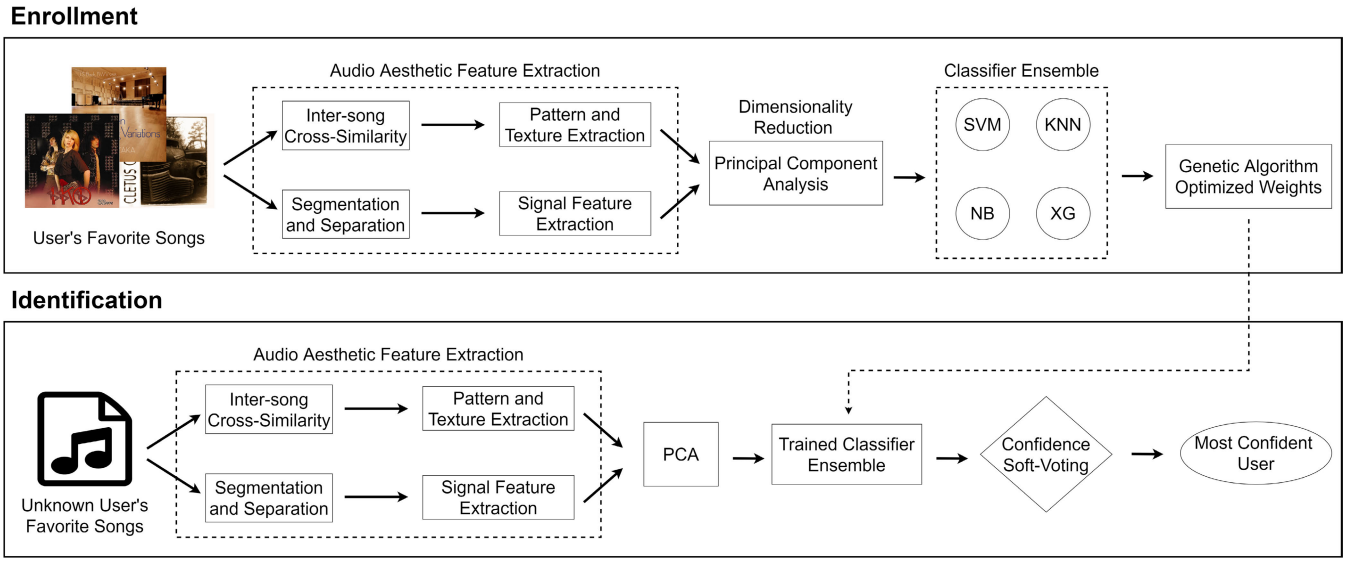


FIGURE 1. A high-level process diagram of the proposed identification system.

are automatically generated, feature representation learning is an important process [19]. The most recent work in aesthetic identification by Bari *et al.* [4] used a custom CNN architecture to analyze user images. A pre-trained network using VGG16 architecture was used for feature extraction and then applied to PCA, obtaining a low-dimensional feature vector with high variance. By doing so, a high-level discriminative feature representation from the images was obtained. Then, an original residual learning-based CNN was used to test and train on the feature representation to obtain a rank 1 accuracy of 97.7%.

Research into audio preferences is a large domain spanning multiple fields of expertise. As a psychological-based concept, some papers explore the properties of musical preference [12] while others investigate its effects on aspects of an individual ranging from personality to substance abuse [20], [21]. The development of music recommendation systems that train on a user's song library to recommend new songs is a well-researched problem in both academia and industry [22]. Kaur, Singh, and Roy published a study on utilizing a person's electroencephalography (EEG) signals when listening to music for the purpose of identification [23]. Using a Hidden Markov Model (HMM) and a dataset of 2400 EEG signals from 60 users, an accuracy of 97.5% was obtained. However, the listening behavior of the users was not restricted to preference in the music, as all EEG waves emitted during the experience were analyzed. In a 2020 IEEE access publication, a Guided Adversarial Autoencoder (GAAE) was used to produce effective learning sample representations from limited audio data [8]. The resulting Inception Score (IS) and Frechet Inception Distance (FID) outperformed other generative models, reinforcing the importance of data representation in audio machine learning models.

The prior research in aesthetic identification has proved that high discriminatory value can be extracted from images liked by users in the visual domain. Feature engineering, along with sophisticated machine learning approaches are both crucial trends in the extraction and usage of these features. The features used in the previous works also only observe the sample level features, but do not consider the explicit relationships between the samples in the set. In addition, there have yet to be any studies exploring whether similar levels of success can be found in the audio domain. This leads to the motivation for this work, and the proposal of the content-based person identification system using audio aesthetics.

III. METHODOLOGY

A. OVERVIEW

The proposed system is an audio aesthetic-based identification system that classifies users from a sample containing audio preferences in music. The samples are sets of songs that the user likes. Given a set of songs, the feature vector is extracted from each song's spectral signal and concatenated to form an audio preference representation. There are two categories of features: inter-song features and intra-song features. Inter-song features relate to the relationship between songs, while intra-song features represent properties of the songs themselves. After extraction, feature selection is used to determine the most discriminatory features from the concatenated original set for identification. The final samples are further divided into 3-song training and testing sets to be analyzed by the ensemble classifier. In order to ensure there are no duplicates and that song order is not learned, combinations are generated rather than permutations for a given observation. Fivefold cross-validation using stratified sampling is used. An abstract view of the procedure is shown in Figure 1.

Recently developed behavioral biometric systems are still predominantly based on classical machine learning

methods [24], [25]. The proposed audio aesthetic system leverages this approach as it is computationally inexpensive, highly effective, and generalizable to other domains.

B. BASELINE DATASET

For the purpose of this research, a dataset of 34 users was collected, where each user was asked to choose ten favorite songs from the set of 224 songs. The 224 songs were sampled from the publicly available Free Music Archive (FMA) dataset for music analysis, while retaining a balanced selection of songs from the genres of Pop, Rock, Folk, Hip-Hop, Jazz, Country, Classical, and Disco. The original FMA dataset consists of 917 gigabytes of Creative Commons-licensed tracks and 161 genres [26]. The medium-sized variant of the FMA dataset was sampled from, which consists of 25,000 tracks of 30-second length and 16 balanced genres. The song clips are played for the user through random order in a controlled environment within a single session.

C. FEATURE EXTRACTION

The features are divided into two categories: intra-song features and inter-song features. Intra-song features describe the individual aspects of each song within the set of liked songs of a user. This includes the aggregated standard deviation and mean of the chromagram features, spectral features, and Mel-Frequency Cepstral Coefficients (MFCC) across frames. The intra-song features intuitively represent the individual structure of each song. Inter-song features describe the relationship between songs within the set. In order to represent this relationship, an affinity cross-similarity matrix is computed and analyzed for each song pair. The inter-song features represent the distance between songs, and in extension, the range of a user's aesthetic preference. If the similarity matrices within the set deviate greatly from the learned matrices, the classifier would be able to distinguish the difference more prominently.

After the extraction of both the intra-song and inter-song features, the two feature categories are concatenated to form the complete feature vector for each sample song set. The training of these samples forms the basis of a user's aesthetic preference template.

1) INTRA-SONG FEATURES

During the feature extraction process, a particular song is separated into segments of even length seconds, which must be factors of 30 seconds. Due to the sampling rate of 22050 Hz, the loaded signal consists of 661500 (22050 x 30) frames. These signals are further separated into component waveforms: the harmonic waveform and the percussive waveform, as shown in Figure 2. The harmonic waveform captures the pitched instruments, and the percussive waveform captures beat/rhythm. A median-based separation technique is utilized that separates the original waveform based on the assumption that harmonic components typically exhibit horizontal patterns while percussive components exhibit vertical patterns on the spectrogram of an audio signal [27], [28]. Harmonic Percussive Source Separation (HPSS) produces more discriminative signals for music classification [29], [30].

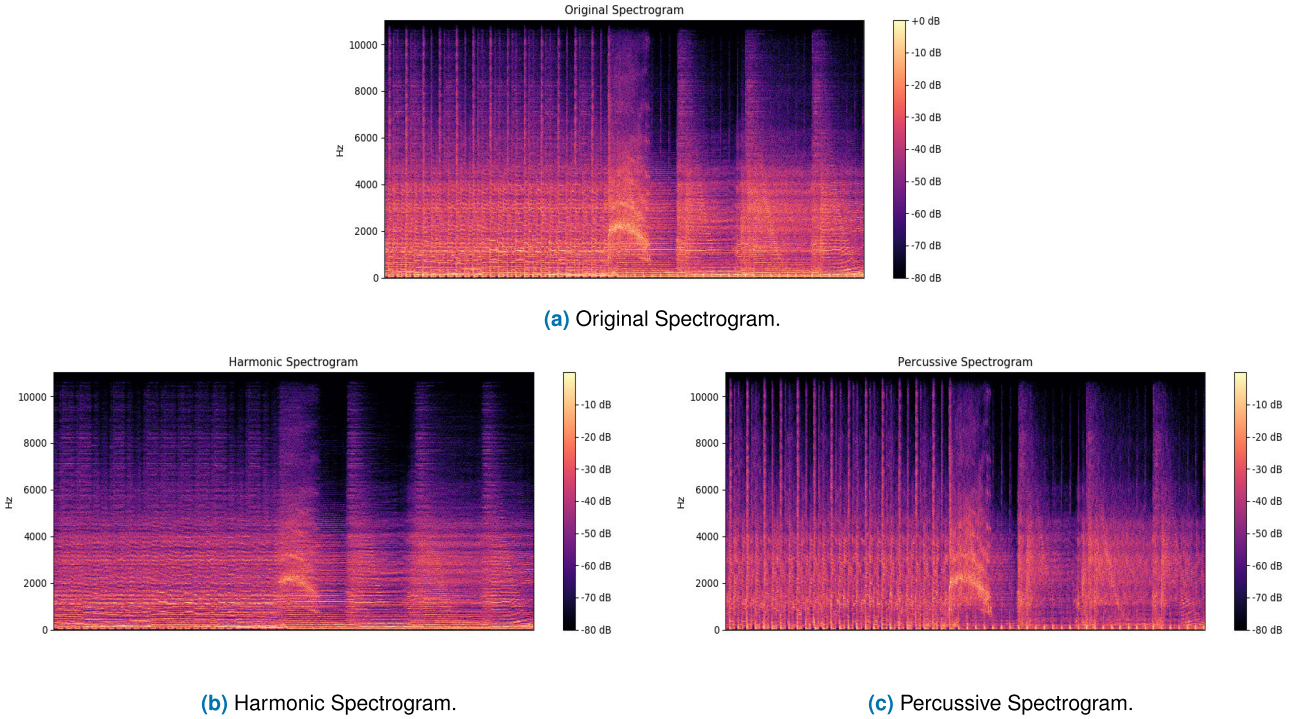
Once the song is split into separate waveform segments, the audio features are extracted for each of these segments. Each song is additionally divided into 3 temporal segments, thus reducing the information loss from signal aggregation. The audio features are then concatenated to form the base feature representation for the song. Each song in the 3-song list possesses a feature representation. To form a representation for a list of songs, the audio features extracted for all the songs are concatenated horizontally to assemble the preference sample. A preference sample corresponds to a set of songs that a particular user likes, which serves as the input to the training and testing sets during the classification phase. A figure of the system diagram is shown in Figure 3.

11 distinct audio features are extracted from a particular segment at a sampling rate of 22050 Hz using the publicly available Librosa audio analysis library. The spectral audio features are extracted as an array of values for each sampled point in time over the audio segment. The mean, standard deviation, minimum and maximum values of each spectral feature array are used to aggregate the array, for a total of 44 features per waveform. The features are common in music recognition and have been used in previous works for similar domains. A small description of the features used is shown in Table 1.

2) INTER-SONG FEATURES

The objective of the inter-song features is to capture the variance in the similarity between song pairs in the set. Thus, a similarity matrix is calculated per song pair. This cross-similarity is a recurrence matrix between two different songs based on affinity and cosine distance along all frames. Unlike the intra-song features, the inter-song features are not segmented by waveform or time windows. In order to produce a more discriminatory cross-similarity matrix, both song signals use short-term history embedding, which is a form of data augmentation that vertically concatenates the signal with past states of itself. A delay of 3 frames and 10 stack dimensions was empirically found to be effective without significant signal degradation and computation costs. The calculated cross-similarity matrices are then saved with a Dots Per Inch (DPI) of 200 and a resolution of 992×739. Each matrix is approximately 150 KB in size. The higher resolution and DPI of the image is important to ensure the image processing of the graph receives as little abstracted data as necessary.

After computing the cross-similarity matrix for each song pair, the pattern and texture features must be extracted in a form that can be combined with the intra-song features. To do this, the Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP) features are computed from the image of the cross-similarity matrices of the song pairs in each set. For the HOG features, 9 histogram bins are used, with a 6×6 sliding window across the image and 1×1 cell blocks. The 9 histogram bins store the directional pattern information of



(a) Original Spectrogram.

(b) Harmonic Spectrogram.

(c) Percussive Spectrogram.

FIGURE 2. A spectrogram comparison for a song. The harmonic spectrogram retains horizontal patterns from the original, while the percussive spectrogram retains the continual vertical beats of the song.

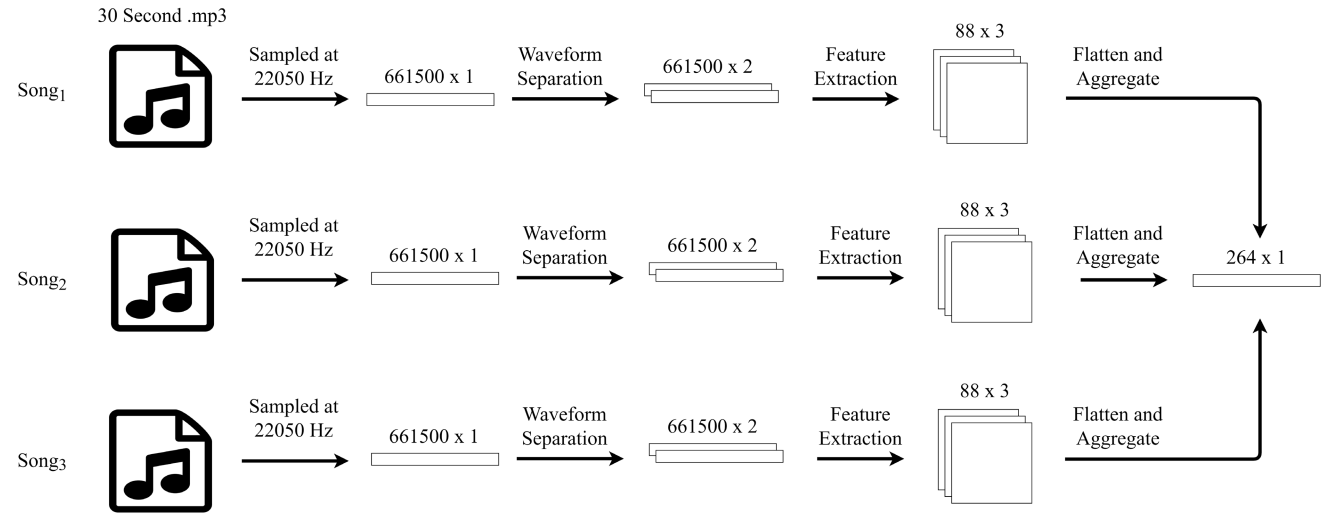


FIGURE 3. The lower-level system diagram for the intra-song feature extraction module. Each input signal is divided into two waveforms and further divided into three temporal segments before concatenation.

the pixels, while the sliding window denotes the abstraction of surrounding aggregated pixels. This extracts a flattened directional pattern image (or HOG image) from the original cross-similarity map. The above configuration produces a histogram of 324 features.

A grid-based LBP approach is used to extract the texture information from partitions of the cross-similarity matrix. Intuitively, the algorithm outputs a flattened texture descriptor by comparing local pixel intensities in a designated neighborhood cell. The image is divided into a 3×3 grid, with a uniform and rotation invariant LBP function applied to each cell. A circle radius of 20-pixel units and 50 circularly symmetric neighbor set points are used as LBP parameters. After traversing a cell, 50 bins are produced from each of the 9 cell histograms from the top-left cell to the bottom-right cell. These cell histograms are then concatenated to form a total number of 450 LBP features per image.

Once the HOG and LBP features are calculated, they are concatenated to form the final inter-song feature matrix. This final inter-song matrix is then concatenated to the intra-song features, and PCA is used to compute the principal components for the final feature vector.

TABLE 1. Features used per frame of the audio signal, calculated as a frame-wise array.

Feature	Description
Waveform Chromogram	Projects waveform onto 12 bins representing the 12 musical octaves.
Constant-Q Chromogram	Applies the Constant-Q transform over the chromogram.
Root Mean Square	The root-mean-square value.
Spectral Centroid	Calculates the weighted mean of the frequency signal. [31]
Spectral Bandwidth	Calculates the width of the signal. [31]
Spectral Contrast	Octave-based relative spectral distribution. [32]
Spectral Flatness	Compares spectral power throughout the signal to determine if the sound is flat (or similar to white noise). [33]
Spectral Roll-off	The frequency at which a designated percentile of the frequency is distributed into the same energy bin or below.
Zero-Crossing Rate	The rate of sign changes (crossing of the zero) across the signal.
Tonnetz	Calculates the tonal centroid feature. [34]
MFCC	Calculates the Mel-Frequency Cepstrum Coefficients, derived from the cepstral representation (spectrum of the spectrum).

D. CLASSIFICATION

An ensemble classifier is used for sample classification that is composed of Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive-Bayes (NB), and XG Boost (XG) classifiers. The ensemble uses a soft-voting strategy to determine the final prediction, which is computed from the average confidence probabilities from all component classifiers. Using an ensemble system ensures multiple learners can contribute to the overall accuracy through diversity when compared to one classifier [35]. Configurations for the classifier hyperparameters are found using a grid search.

After the feature vector has been generated for each user, the testing and training sets are generated for the ensemble classifier. $\binom{10}{3}$ unique 3-song sets are generated from the total 10 liked songs of the user that retain the aesthetic preference of the user. This ensures that there are no identical sets within the population and an identical set with different song orders is not counted. Each 3-song set composes one sample of the training or testing process. The mean accuracy of the 5-fold cross-validation is used, with random splitting and shuffling of the test and train indices each fold.

TABLE 2. Hyperparameter configuration for the component classifiers using grid search.

Classifier	Hyperparameters
SVM	0.25 squared l2 regularization, Radial Basis Function (RBF) kernel and a stopping tolerance of 0.001 .
KNN	Uniform weight, 20 neighbors and leaf size of 30.
NB	Variable smoothing of 10^{-8}.
XG	Subsample column ratio of 0.8 for each tree, step size shrinkage of 0.2, minimum child weight of 1, training subsample ratio of 0.5 and the histogram tree method.

As the performance of each individual classifier differs depending on the distribution, the decision weight of each classifier can also be tuned to produce a stronger average ensemble [36]. To tune the decision weights of each component classifier, a genetic algorithm is used. A Genetic Algorithm (GA) is a meta-heuristic optimization technique that leverages evolutionary programming to search for a solution within a solution space. Principles of natural selection and genetic operations are used to maintain a population of potential solutions at one time. Most notably, the mutation operation implements randomness into the population per generation and the crossover operation allows for chromosomes in candidates to transfer to their children. In this problem, the objective function of the genetic algorithm is to minimize the error rate of the 5-fold cross-validation, and the chromosomes are the weights of each component classifier in the ensemble. Each weight is optimized within the real range of [0,1]. A uniform crossover rate of 0.7 is used, which uniformly distributes the variables in the parent chromosomes to the offspring. The crossover rate corresponds to the ratio of offspring in the population which are products of crossover. A mutation rate of 0.3 is used to introduce randomness into the population. Throughout 100 generations with a population of 30, singular elitism is employed and a parent ratio of 0.2 is retained across generations. Singular elitism refers to the retaining of the fittest individual in the population across each generation, while the parent ratio describes the portion of the next generation that is composed of candidates from the previous generation. These hyperparameters were found empirically.

IV. EXPERIMENTAL RESULTS

The performance of the intra-song features and the inter-song features is shown in Figure 4 and Table 5. As a property of features, the low performance of an isolated feature is not

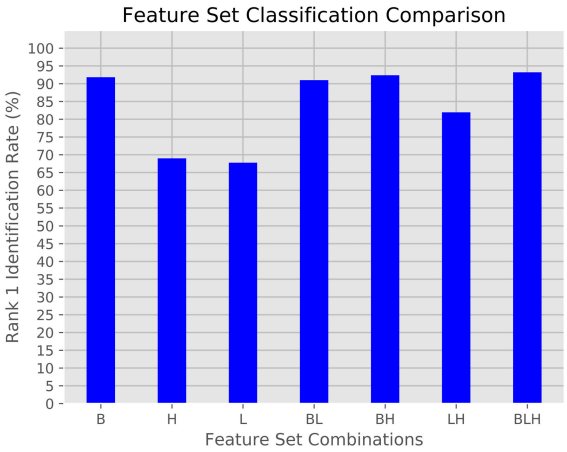


FIGURE 4. A comparison between the different feature sets. B denotes the intra-song features, L denotes the LBP inter-song features and H denotes the HOG inter-song features.

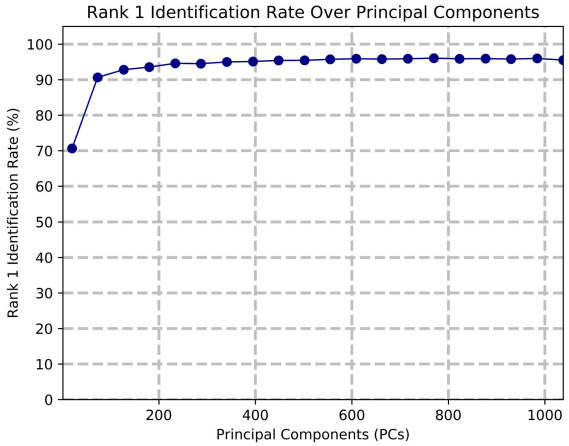


FIGURE 5. A comparison between different principal components on rank 1 identification rate using the ensemble system.

TABLE 3. Comparison of rank 1 accuracy between various feature subsets used in the system.

Feature Set	Rank 1 Accuracy
B	91.81%
H	68.97%
L	67.75%
BL	91.00%
BH	92.38%
LH	81.94%
BLH	**93.21%**

necessarily indicative of its performance within a pool of features. Thus, subsets of the feature pool are tested. B denotes the base features used commonly in music classification problems, H denotes the HOG features and L denotes the LBP features. Results on the dataset show that the B, BL, BH and BLH features perform with over 90% accuracy in an isolated environment. The HOG and LBP feature sets exhibit lower accuracy in isolation, but certain combinations of the features produce increased accuracy in comparison. The performance of the LH and BLH sets shows that the combination of both texture and pattern information leads to higher accuracy. The different combinations of the 3 feature categories are tested, with the BLH set producing the highest accuracy of 93.21%. This can be attributed to the contribution of information among the inter-song (H, L) and intra-song (B) features. The inter-song features are shown to capture additional discriminatory information of the relationship between song pairs that would not be present in only the intra-song features. This combined BLH feature set is used for the final system and subsequent experiments.

Dimensionality reduction is applied for the combined feature set obtained during the previous step. Figure 5 shows the comparison of varying principal components on rank 1 identification accuracy. From the original 1038 feature set, 770 principal components were found to provide the highest increase in rank 1 identification rate. An accuracy of 96.05% is achieved while lowering the feature size by approximately 26%. Less principal components do not possess as much discriminatory information and more principal components introduce noise that degrades performance. Starting at a threshold of 448 principal components, a decrease in principal components produces a loss in performance for the system while an increase causes only mild fluctuations in performance. At this point, an accuracy of 95.42% is achieved. Although the accuracy is not at its highest point, there is only a loss in accuracy of 0.63% while only using 448 principal components when compared to the 770 principal component scenario. This can be preferable when performance in a lightweight system is desired due to a feature size reduction of approximately 57%. 448 principal components are chosen for the resulting system to retain the computational efficiency and scalability of the dataset. The training times of the component classifiers and the ensemble is shown in Table 4.

TABLE 4. Comparison of classifier training times in seconds using 448 principal components.

Classifier	Training Time
SVM	32.34 seconds
KNN	0.053 seconds
NB	0.014 seconds
XG	23.36 seconds
Ensemble	55.24 seconds

The same experiment is run on the component classifiers in isolation with optimized grid search configurations, as shown in Figure 6. Support Vector Machine is denoted by SVM, K-Nearest-Neighbor by KNN, Naive Bayes by NB, and XGBoost by XG. The results show that the data is more easily differentiated by SVM, NB, and XG, with slightly lower

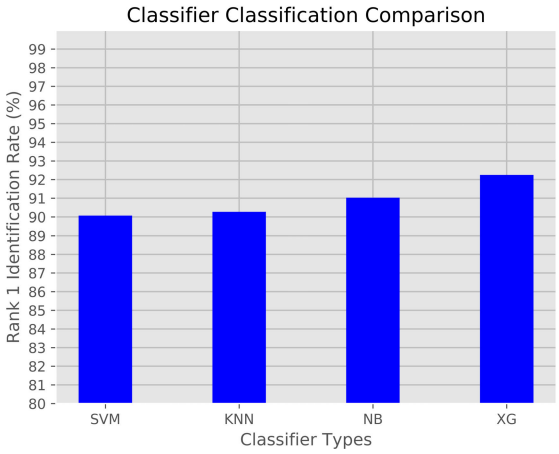


FIGURE 6. Comparison of rank 1 accuracy between the various component classifiers used in the ensemble.

TABLE 5. Tabulated rank 1 accuracy classifier comparisons.

Classifier	Rank 1 Accuracy
SVM	90.07%
KNN	90.27%
NB	91.03%
XG	92.25%

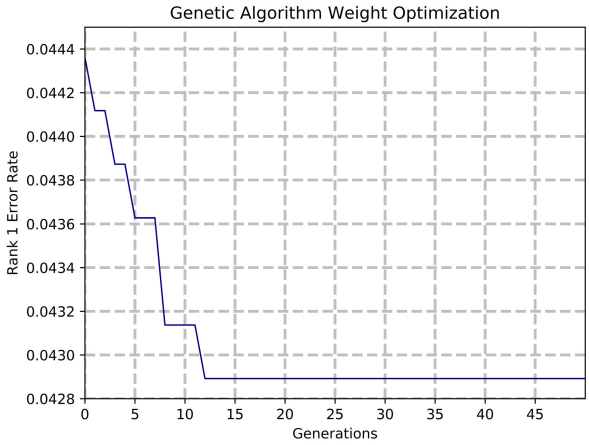


FIGURE 7. The rank 1 identification error-rate curve for the first 50 generations of the genetic algorithm.

accuracy for KNN. All component classifiers exhibit accuracies above 90%, with an average classifier performance of 91.39%. Due to the inclusion of both similarly performing linear and non-linear classifiers, a more diversified decision boundary can be achieved for the final ensemble system.

A genetic algorithm is used to select the weights of each component classifier within the ensemble system, with the results shown in Figure 7. The convergence of the genetic algorithm in minimizing rank 1 identification error is primarily evident in the preliminary iterations. As the optimization solution space is comparatively small with four parameters, a near-optimal solution can be found quickly. The best candidate in the randomly initialized population at generation 1 has an error rate of 4.43%, but this decreases to 4.26% within the first 15 generations. Due to the elitism strategy carrying the best-known candidate from the previous generation to the next, the best solution is always propagated forward. This results in consistent accuracy after generation 15, as the solution space is being explored with no significantly better candidate found for up to 100 generations (only the first 50 generations are shown in the graph). The final accuracy of the ensemble obtained through weight optimization using GA is 95.74%. This configuration of optimized weights is used in the final system and all subsequent experiments.

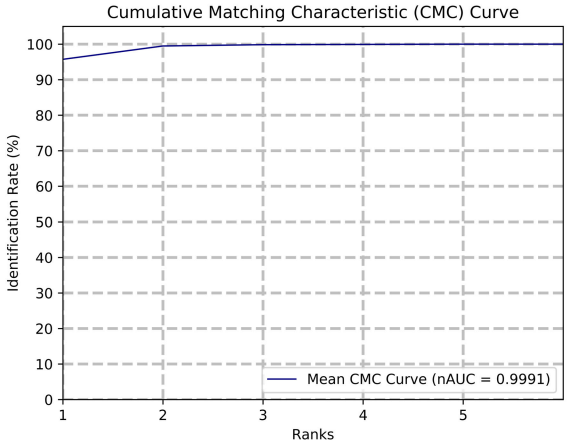


FIGURE 8. The Cumulative Matching Characteristic (CMC) curve showing the rank n identification for the first 6 ranks.

The Cumulative Matching Characteristic (CMC) curve is shown in Figure 8, which displays the proposed classification system's accuracy across rank 1 to rank 5 recognition rates. The CMC curve in a person identification system represents the system's reliability in correctly identifying users within a number of predictions. For example, a rank 1 identification rate is the rate at which the system correctly predicts the user in one prediction, while a rank 5 identification rate would be the rate of a correct prediction within the top five predictions. The normalized Area-Under-the-Curve (nAUC) of a CMC curve is a unit of measure for the overall accuracy of the CMC curve, where a completely accurate system would have an ideal nAUC of 1. The system achieves a normalized AUC of 0.9991 among all 34 user classes, with a rank 1 recognition of 95.74% and a rank 5 recognition of 100%.

The Receiver Operating Characteristic (ROC) curve in Figure 9 shows the relation between the system's True Positive Rate (TPR) over the False-Positive Rate (FPR). A high TPR followed by a low FPR shows that the system has fewer verification errors. Similar to the nAUC in the CMC curve, the area-under-the-curve of the ROC curve also is indicative of the accuracy of the system and tends to an ideal value of 1. From this information, the False-Negative-Rate (FNR) and the False-Positive-Rate (FPR) can be determined.

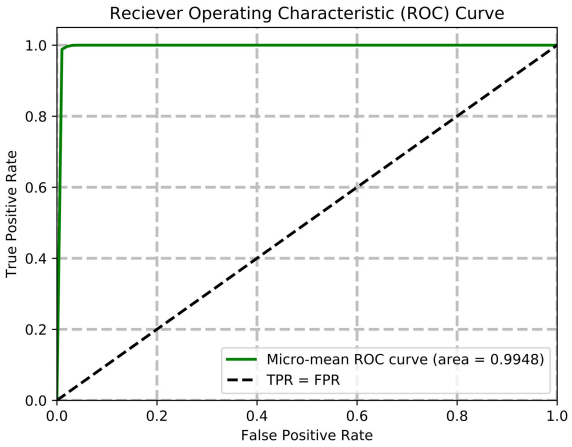


FIGURE 9. The Receiver Operating Characteristic (ROC) curve showing the system's true positive rate over false positive rate.

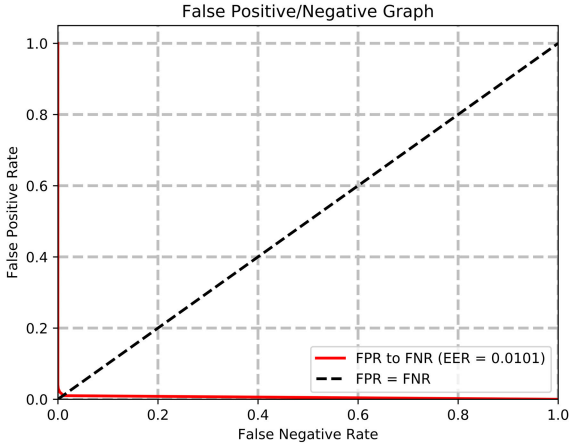


FIGURE 10. The error rate curve showing the system's false positive rate over false negative rate and Equal Error Rate (EER).

The false-positive/negative graph in Figure 10 shows the FPR (Type I error) over the FNR (Type II error). The Equal Error Rate (EER) is the threshold at which the two error rates are equal. A false positive represents the acceptance of an incorrect user, and a false negative represents the incorrect rejection of a correct user. The equal error rate is the point at which the false positive rate and the false-negative rate intercept. A low EER is desired for a biometric system, as this indicates it is less prone to both types of verification error. In this identification problem, the multiclass micro-mean is recorded, where the ROC and EER metrics are the averages of all the binary class scenarios using a One-Versus-Rest (OVR) approach. The model exhibits a ROC area-under-the-curve of 0.9948, with an equal error rate of 0.0101.

To demonstrate the system's portability, another dataset is constructed using the Million Playlist Dataset (MPD) from Spotify [37]. 200 playlists from anonymous Spotify users are sampled from the first 1000 MPD playlists containing 10 songs each. 30-second clips of each song are downloaded using the Spotify Web API to allow for content-based feature

TABLE 6. Comparison of rank 1 and 2 accuracies and inference times between the Free Music Archive (FMA) and Million Playlist Dataset (MPD) datasets.

Dataset	Rank 1 Acc.	Rank 2 Acc.	Inference Time
FMA	95.74%	99.49%	1.85 s
MPD	99.60%	99.98%	8.12 s

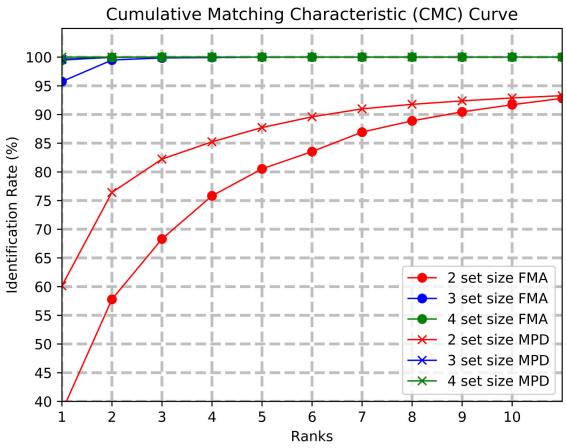


FIGURE 11. CMC curve for song sets containing 2 to 5 user songs for both the FMA and MPD datasets.

extraction, mirroring the dataset collected in this work. Due to the unavailability of certain song clips or insufficient playlist lengths, unviable samples are discarded. An identical procedure of pre-processing and sampling is used as the collected FMA dataset. A comparison of the inference time taken per user and rank 1/2 accuracies of the system are shown in Table 6. Despite the higher user count, the system still performs with high accuracy by exhibiting above 95% rank 1 accuracy. Songs in the FMA dataset are sampled from a reduced pool of 224 songs, while the songs in the MPD dataset have no limitation. Thus, the increase in accuracy for the MPD dataset over the FMA dataset can be attributed to a larger song diversity, resulting in more discriminative extracted features. The difference in inference time of the two datasets is anticipated due to the disparity in user count.

Using these hyperparameters, the performance of the system under varying song set sizes is compared in Figure 11. In a practical implementation, aspects including data quality, availability, and integrity can produce sparse samples, which can affect the performance of a system. Intuitively, the performance of the system can be expected to decrease as less viable information is made available. Insufficient data can lead to a classifier's incorrect learning and ultimately, prediction [38]. This is shown primarily in the 2-song set experiments, which exhibit lower rank 1 accuracies in comparison to higher song set counts. At a baseline of 3-song set size, rank 1 accuracy is shown to increase dramatically until consecutive increases in set size plateau at approximately 100%. A 3-song set size maintains significant accuracy over a 2-song set and requires less information from the user than a 4-song set. This

TABLE 7. Rank 1 accuracy and inference time for visual, audio and combined aesthetic identification systems.

System	Accuracy	Inference Time
Visual (Bari et al. [4])	98.38%	1.98 s
Proposed Audio	95.74%	1.85 s
Proposed Combined	99.41%	2.84 s

lowered requirement allows for more flexibility when data is limited, while also reducing feature processing time and storage needed.

Through the series of experiments, the discriminatory value of the base, LBP, and HOG features are shown. 448 principal components of the base, LBP and HOG feature sets are then found to produce a high accuracy of 93.21%. The final system proves to be accurate with a CMC nAUC of 0.9991 and robust to error with a ROC AUC of 0.9948. The system is then experimented on two constructed datasets, both with comparatively high performance. As this is the first proof-of-concept work that establishes the possibility of user identification using audio aesthetic preferences, no prior research in the audio aesthetics domain can be compared to the reported performance. However, Table 7 lists the inference time and the recognition accuracy of a visual aesthetic-based system for the same set of users for which the audio set was collected. The visual aesthetic system was tested on images taken from Flickr, while the audio aesthetic system was tested on the FMA dataset.

V. CONCLUSION AND FUTURE WORK

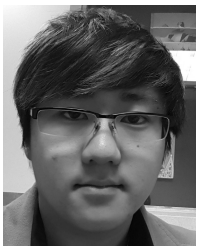
This paper introduces the first system for person identification using audio aesthetics. This system has been developed and tested on two newly constructed audio aesthetic datasets. The research establishes that audio features possess unique characteristics that allow for accurate user recognition. The system achieves a rank 1 accuracy of 95.74% with 448 principal components by utilizing both inter-song and intra-song features. The feature combinations are tested, with a combination of base song, LBP, and HOG features yielding the highest accuracy. Within a pool of users, it is possible for a system to accurately differentiate specific users given the appropriate features. This shows that audio aesthetic features hold discriminatory value similar to the visual counterpart. Furthermore, a score-level fusion of the proposed audio system and the visual system achieves 99.41% rank 1 accuracy. As person identification using audio aesthetic features is a new domain, these results serve as a proof-of-concept for audio aesthetic features, with the constructed datasets facilitating further research.

In the future, the exploration of deep learning approaches may prove insightful, especially due to the recent success of convolutional neural networks in visual aesthetic identification. In addition to exploring deep learning methods for the audio-based aesthetic system, another potential direction is to investigate the correlation between audio and image-based aesthetic choices of the users. The investigation of the effect of various fusion methodologies on the overall performance of the combined audio and visual system is a promising avenue as well.

REFERENCES

[1] M. Sultana, P. P. Paul, and M. L. Gavrilova, "Social behavioral information fusion in multimodal biometrics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2176–2187, Dec. 2018.
[2] P. O. Kristeller, "The modern system of the arts: A study in the history of aesthetics Part I," *J. Hist. Ideas*, vol. 4, pp. 496–527, Oct. 1951.
[3] M. Moshagen and M. T. Thielsch, "Facets of visual aesthetics," *Int. J. Hum.-Comput. Stud.*, vol. 68, no. 10, pp. 689–709, 2010.
[4] A. H. Bari, B. Sieu, and M. L. Gavrilova, "Aestheticnet: Deep convolutional neural network for person identification from visual aesthetic," *Vis. Comput.*, vol. 36, no. 10, pp. 2395–2405, 2020.
[5] P. Lovato, A. Perina, N. Sebe, O. Zandonà, A. Montagnini, M. Bicego, and M. Cristani, "Tell me what you like and I'll tell you what you are: Discriminating visual preferences on Flickr data," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, p. 4556.
[6] S. Azam and M. Gavrilova, "Gender prediction using individual perceptual image aesthetics," *J. WSCG*, vol. 24, no. 2, pp. 53–62, 2016.
[7] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58043–58055, 2018.
[8] K. N. Haque, R. Rana, and B. W. Schuller, "High-fidelity audio generation and representation learning with guided adversarial autoencoder," *IEEE Access*, vol. 8, pp. 223509–223528, 2020.
[9] W. W. Y. Ng, W. Zeng, and T. Wang, "Multi-level local feature coding fusion for music genre recognition," *IEEE Access*, vol. 8, pp. 152713–152727, 2020.
[10] R. Yang, L. Feng, H. Wang, J. Yao, and S. Luo, "Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices," *IEEE Access*, vol. 8, pp. 19629–19637, 2020.
[11] M. Williams, "Preference for popular and world music: A review of literature," *Update, Appl. Res. Music Educ.*, vol. 35, no. 3, pp. 31–37, Jun. 2017.
[12] T. Schäfer and P. Sedlmeier, "What makes us like music? Determinants of music preference," *Psychol. Aesthetics, Creativity, Arts*, vol. 4, no. 4, p. 223, 2010.
[13] E. J. Vella and G. Mills, "Personality, uses of music, and music preference: The influence of openness to experience and extraversion," *Psychol. Music*, vol. 45, no. 3, pp. 338–354, May 2017.
[14] C. Segalin, A. Perina, and M. Cristani, "Personal aesthetics for soft biometrics: A generative multi-resolution approach," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 180–187.
[15] S. Azam and M. Gavrilova, "Person identification using discriminative visual aesthetic," in *Proc. Can. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2017, p. 1526.
[16] B. Sieu and M. Gavrilova, "Biometric identification from human aesthetic preferences," *Sensors*, vol. 20, no. 4, p. 1133, Feb. 2020.
[17] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, vol. 7, pp. 41273–41285, 2019.
[18] S. Jang, L. Battulga, and A. Nasridinov, "Detection of dangerous situations using deep learning model with relational inference," *J. Multimedia Inf. Syst.*, vol. 7, no. 3, pp. 205–214, Sep. 2020.
[19] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.
[20] J. Mulder, T. F. M. Ter Bogt, Q. A. W. Raaijmakers, S. N. Gabhainn, K. Monshouwer, and W. A. M. Vollebergh, "The soundtrack of substance use: Music preference and adolescent smoking and drinking," *Substance Use Misuse*, vol. 44, no. 4, pp. 514–531, Jan. 2009.
[21] P. G. Dunn, B. de Ruyter, and D. G. Bouwhuis, "Toward a better understanding of the relation between music preference, listening behavior, and personality," *Psychol. Music*, vol. 40, no. 4, pp. 411–428, Jul. 2012.

[22] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *Proc. 9th Int. Symp. Comput. Music Modeling Retr.*, vol. 4, 2012, pp. 395–410.
[23] B. Kaur, D. Singh, and P. P. Roy, "A novel framework of EEG-based user identification by analyzing music-listening behavior," *Multimedia Tools Appl.*, vol. 76, no. 24, pp. 25581–25602, Dec. 2017.
[24] S. Dargan and M. Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities," *Expert Syst. Appl.*, vol. 143, Apr. 2020, Art. no. 113114.
[25] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," 2019, *arXiv:1912.00271*. [Online]. Available: http://arxiv.org/abs/1912.00271
[26] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proc. 18th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2017.
[27] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. DAFX*, vol. 10, 2010, pp. 1–4.
[28] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *Proc. ISMIR*, 2014, pp. 611–616.
[29] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagayama, "Autoregressive MFCC models for genre classification improved by harmonic-percussion separation," in *Proc. ISMIR*, 2010, pp. 87–92.
[30] A. Rosner, B. Schuller, and B. Kostek, "Classification of music genres based on music separation into harmonic and drum components," *Arch. Acoust.*, vol. 39, no. 4, pp. 629–638, Mar. 2015.
[31] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Berlin, Germany: Springer, 2007.
[32] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2002, pp. 113–116.
[33] S. Dubnov, "Generalization of spectral flatness measure for non-Gaussian linear processes," *IEEE Signal Process. Lett.*, vol. 11, no. 8, pp. 698–701, Aug. 2004.
[34] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. 1st ACM workshop Audio Music Comput. Multimedia*, 2006, pp. 21–26.
[35] A. Rahman and S. Tasnim, "Ensemble classifiers and their applications: A review," 2014, *arXiv:1404.4088*. [Online]. Available: http://arxiv.org/abs/1404.4088
[36] A. Ekbal and S. Saha, "Weighted vote-based classifier ensemble for named entity recognition: A genetic algorithm-based approach," *ACM Trans. Asian Language Inf. Process.* , vol. 10, no. 2, pp. 1–37, Jun. 2011.
[37] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani, "Recsys challenge 2018: Automatic music playlist continuation," in *Proc. 12th ACM Conf. Recommender Syst.*, Sep. 2018, pp. 527–528.
[38] R. Blake and P. Mangiameli, "The effects and interactions of data quality and problem complexity on classification," *J. Data Inf. Qual.*, vol. 2, no. 2, pp. 1–28, Feb. 2011.



BRANDON SIEU (Member, IEEE) received both the B.Sc. degree in computer science and the B.Comm. degree in business technology management from the University of Calgary, in 2019, where he is currently pursuing the M.Sc. degree. He published research in the International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC) 2018/2019, Computer Graphics International (CGI) 2019, and CyberWorlds (CW) 2019. His research interests include biometrics, pattern recognition, and visual/audio aesthetics. In addition, he has served as a referee for the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS (TCSS) journal.

MARINA L. GAVRILOVA (Senior Member, IEEE) is currently a Full Professor with the Department of Computer Science, University of Calgary, where she is also the Head of the Biometric Technologies Laboratory. Her publications include over 200 journal articles and conference papers, edited special issues, books, and book chapters in the areas of image processing, pattern recognition, machine learning, biometric, and online security. She is a Board Member of the ISPIA. She has given over 50 keynotes, invited lectures, and tutorials at major scientific gatherings and industry research centers, including at Stanford University, SERIAS Center at Purdue, Microsoft Research, USA, Oxford University, U.K., and Samsung Research, South Korea. She has founded ICCSA–an international conference series with LNCS/IEEE, co-chaired a number of top international conferences, and is the Founding Editor-in-Chief of *LNCS Transactions on Computational Science* journal. She currently serves as an Associate Editor for IEEE ACCESS, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, *The Visual Computer*, and the *International Journal of Biometrics*, and was appointed by the IEEE Biometric Council to serve on IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE Committee.