IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 July 2022, accepted 10 August 2022, date of publication 18 August 2022, date of current version 31 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3200166

RESEARCH ARTICLE

A Novel Three Stage Framework for Person Identification From Audio Aesthetic

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

Corresponding author: Fariha Iffath (fariha.iffath@ucalgary.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, in part by the NSERC Strategic Partnership Grant (SPG), and in part by the Innovation for Defense Excellence and Security Network (IDEaS).

ABSTRACT Social behavioral biometrics investigates social interactions to determine a person's identity. Within the discipline of social behavioral biometrics, recognition of individuals based on their aesthetic preferences is an emerging direction of research. Human aesthetic is a soft, behavioral biometric trait that refers to a person's attitudes towards a particular subject material. Recent developments in aesthetic-based biometric systems have proven that an individual's visual and audio aesthetic preferences hold considerable distinctive features. This paper introduces a novel three-stage audio-aesthetic system that can uniquely identify a user from the set of their favorite songs. The system utilizes Residual Network (ResNet) for high-level feature extraction. A hybrid meta-heuristic feature selection algorithm based on Cuckoo Search and Whale Optimization is proposed for feature extraction optimization, which results in the low-dimensional feature set. The selected subset of features is fed into the XGBoost classifier to establish a person's identity. The proposed method outperformed the handcrafted feature-based method by achieving 99.54% accuracy on a proprietary dataset (Free Music Archive) and 99.79% accuracy on a publicly available dataset (Million Playlists Dataset).

INDEX TERMS Social behavioral biometrics, deep learning, biometric authentication, audio aesthetics, transfer learning, meta-heuristic, feature selection.

I. INTRODUCTION

Biometric systems establish identity of an individual based on unique physical or behavioral attributes. Physiological biometric systems encode an individual's physical characteristics to create a template unique to that person. Most commonly used physiological biometrics are fingerprints, iris, palm, face, and hand geometry [1]. On the other hand, behavioral biometric modalities focus on a person's actions, such as voice, signature, or gait rather than their physical characteristics. Within this domain, social-behavioral biometric investigates a person's identity using their social interactions and communication patterns [2]. Over the last few decades, with the flourishment of technology, the number of social media users has exploded. As a result, social network platforms have become a widespread source of data that can be utilized to verify a user remotely and covertly.

Exploiting personal aesthetic properties for biometric identification is an emerging research direction in social behavioral biometrics. Aesthetic features refer to an individual's preference toward a particular subject material. Several studies have shown that an individual's aesthetic attributes can be utilized to differentiate them from others [3], [4]. With the exponential growth of online social media users, aesthetic data is becoming more ubiquitous in the form of text, photographs, videos, and music. Additionally, people can publicly express their preferences and opinions on different social media platforms, increasing the rapid accessibility of aesthetic data. Furthermore, studies of cognitive neuroscience have emphasized that there is a significant link between human perceptions of aesthetics and personality, making personal aesthetics one of the prominent and desirable traits for biometric authentication [4]. Aesthetic biometric systems

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

have potential to transform many existing applications, such as multi-factor authentication, human behavioral analysis, and recommender systems [5].

In the domain of social behavioral biometric research, aesthetic systems have shown great potential. Several visual aesthetic systems have been recently developed for person identification [6] and gender prediction [7]. These systems have demonstrated that a user's preferred set of images can hold discriminatory features. Aesthetic data is easily accessible and retrievable from online social media platforms. Moreover, combination of aesthetic features with other adaptive biometric traits has a wide range of applications. Such systems can be extended to understand consumer behavior and experience, optimize the positioning of an e-commerce platform, increase the acquisition of new products, and enhance collaborative learning experience [8].

Recent studies have demonstrated the potential applications of audio systems in security and surveillance. Some of the related works include detection of the sound-based drones to protect the security-sensitive institutions [9], prediction of crime by incorporating audio modality with text sentiment [10], and the surveillance of streets based on atypical sounds [11].

The very first audio aesthetic system was developed in 2021 and achieved a significant accuracy by extracting handcrafted audio features from user liked-songs [5]. This research demonstrated that, similar to visual aesthetics, audio aesthetic features have unique and distinguishing characteristics for biometric identification.

Previous audio aesthetic-based person identification research only explored traditional machine learning approaches, trained with handcrafted features. However, extracting and optimizing handcrafted features requires extensive feature engineering, leading to increased computational complexity. In addition, when the size of the dataset increases, the system might face issues such as scalability, reliability, and robustness. On the other hand, Deep Learning (DL) approaches have proven to be highly useful in various domains, including audio analysis, text analysis, and image processing [12]. Deep learning methods can extract features automatically from a vast amount of data [13]. This property can be leveraged to extract high-dimensional features without extensive feature engineering. As a result, this will resolve the scalability and reliability issues of handcrafted feature based systems.

In this work, the following research questions will be addressed:

1) Which deep learning based pre-trained architecture is most suitable for user recognition using audio aesthetic features?
2) Will the high-level features extracted from a pre-trained Convolutional Neural Network (CNN) be more discriminative than handcrafted features?
3) Which hybrid meta-heuristic feature selection algorithm can select the most discriminating audio aesthetic feature subset?
4) Can a hybrid meta-heuristic feature selection algorithm choose a reduced subset of high-level features to achieve a high classification accuracy?
5) Which machine learning classifier can identify users from their preferred set of music with the highest accuracy?

This paper proposes a novel three-stage deep learning based audio aesthetic system dedicated to person identification. To the best of our knowledge, this is the first audio aesthetic system that utilizes deep learning to extract aesthetic features. Initially, the proposed method will extract high-level features using pre-trained deep learning architecture. Later, a hybrid meta-heuristic feature selection algorithm will be employed for selecting the most optimal feature set, which will be further trained using machine learning algorithms for classification. It will be demonstrated that the proposed architecture surpasses the existing audio aesthetic method on two benchmark datasets.

This paper makes the following contributions to answer the research questions:

- A novel three-stage framework based on deep learning and classical machine learning is proposed for identifying individuals from their audio aesthetic.
- To extract audio features, mel-spectrograms are used instead of generic spectrograms. As a result, the extracted feature set becomes more discriminating for person identification.
- Residual network is employed for extracting the most optimal high-level features from users' preferred set of music.
- A novel hybrid meta-heuristic algorithm Cuckoo Search based Whale Optimization Algorithm (CSWOA) is proposed that retrieves the most optimal feature subset from the high-level features.
- Reduced feature set is trained using eXtreme gradient Boosting (XGBoost) classifier for identification in order to reduce computational time.
- The high-level feature extractor ResNet is compared with InceptionNet and VGG16 to prove its superiority.

Extensive comparison of the proposed hybrid meta-heuristic feature selection algorithm with other standalone and hybrid wrapper-based feature-selection methods demonstrates its superiority for selecting audio-based aesthetic feature. The findings prove that the proposed architecture can extract a more discriminating feature set than the previously used handcrafted features. In addition, the experiments show that the proposed approach is more efficient, rigorous, and requires less computational cost than the most recent method for person identification using audio aesthetics. The proposed system attains overall recognition accuracy of 99.54% on the Free Music Archive (FMA) dataset and 99.79% accuracy on the Million Playlist Dataset (MPD) dataset, outperforming state-of-the-art method for audio aesthetic-based person identification.

The remainder of the paper is organized as follows: a detailed overview of existing research on aesthetic-based

systems will be discussed in Section II. The proposed method will be thoroughly described in Section III. Experimental results and comparison of the proposed method with the previous works will be discussed in section IV. Finally, future research directions will be outlined in section V.

II. LITERATURE REVIEW

Recent progress in social behavioral biometric research has introduced various avenues to investigate human behavior based on social media activities, communications, and interactions. Social media became a platform for sharing personal interests, lifestyles, and ideas with acquaintances. As a result, online social media platforms have emerged as a prominent source of information about an individual's behavioral traits. Initially, analysis of this information was limited to linguistic authorship recognition [14], authentication via social media interactions, and spatio-temporal information mining. However, authentication based on human aesthetic preferences remained largely unexplored.

The intuition behind aesthetic-based identification is that every person has distinctive tastes and preferences when it comes to photographs, arts, music, and so on. Such idiosyncratic attributes can be exploited to distinguish individuals from others. The first concept of person identification from their visual preferences was introduced by Lovato *et al.* [15]. In this work, the dataset was sampled from the extensive database of the Flickr website containing 200 users, where each user was asked to select 200 favorite images. A machine learning method named Least Absolute Shrinkage and Selection Operator (LASSO) regression was exploited to learn the most distinguishing aesthetic features, yielding rank 1 identification accuracy of 14%. Despite the low recognition accuracy, this work unveiled a promising research scope for utilizing human preferences as a unique identifier. Later, the same group of researchers carried out another experiment on the Flickr dataset [3], this time integrating new features. This work achieved a rank 1 accuracy of 76%. At the same time, Segalin *et al.* [4] adopted the counting grid model and support vector machine as a learning method instead of LASSO regression utilizing the same database as [15]. In this work, a generative embedding strategy was followed by considering Bags of Features with 111 features. As a generative step, a multi-resolution counting grid was utilized for generating an ensemble of embedding maps. The SVM classifier was trained in a one-versus-all modality achieving rank 1 accuracy of 73%. The major limitation of this approach was that the same image could not be chosen by more than one user. In 2016, Azam and Gavrilova [7] introduced the proof of concept of gender identification of an individual using the perceptual aesthetic features of their preferred images. In this work, the authors utilized a bag of 56 perceptual image aesthetic features. For final classification, the decisions of three traditional binary classifiers were combined using feature selection and ensemble weight adjustment methods. The model was trained and evaluated on a database of 24,000 images from 120 Flickr users, achieving 77% rank 1 accuracy in aesthetic based gender prediction.

The most recent work exploited the original deep learning approach [16] for the first time on visual aesthetic-based identification [6]. Above mentioned techniques were highly dependent on manual feature engineering. However, deep learning models are capable of executing feature engineering on their own [13]. In addition, recent breakthroughs of deep learning models in the domain of computer vision [17] and biometric system [18] led to deployment of a deep learning approach for visual aesthetic systems. An original framework, AestheticNet was developed, yielding 97.7% rank 1 identification accuracy. A pre-trained VGG16 network was used for extracting high dimensional feature maps, which were further reduced to low-dimensional feature vectors with high variance using Principle Component Analysis (PCA). Subsequently, a residual learning-based CNN was used to train and validate the obtained low dimensional feature vector.

Audio preferences have become a ubiquitous phenomenon, which has been explored in multiple areas ranging from psychological research to medical therapy [19]. Apart from visual aesthetics, personal audio preferences have also shown great potential in analyzing an individual's behavioral and psychological traits [20]. Several music recommendation systems have been developed to exhibit the relationship between a user's music preferences and personality type [21], [22]. In addition, the development of human cognitive-based authentication system while listening to music is another prominent research domain. In 2020, Patel and Husain [23] proposed the development of an Electroencephalogram (EEG)-based person authentication system that measured the user's neurophysiological responses while listening to their preferred music. This study aimed at creating a user-authentication system to identify a user uniquely based on their corresponding EEG response to music. Another authentication system, MusicID was developed by utilizing users' preferred set of music [24]. In this work, authors exploited human brainwave patterns while user's listened to their favorite songs.

Prior works have established a correlation between human personality and audio preferences. The first proof of concept research in this domain in 2021 [5]. Authors developed the first audio aesthetic-based person identification system that achieved 95% user recognition accuracy on FMA and MPD datasets. This work utilized intra-song and inter-song features of users' favorite songs. An ensemble classifier was used for final authentication, where each classifier's decision weight was optimized using a genetic algorithm.

According to the prior research, it is evident that, similar to visual aesthetics, a person's audio preferences also hold discriminatory features to identify a user. However, prior works in this domain rely heavily on feature engineering. This can create challenges, including dataset bias, scalability, and reliability issues. Furthermore, the recent success of convolutional neural networks in biometric identification has

demonstrated the efficacy of deep learning architecture in this domain. The above-mentioned points motivate us to develop the first audio aesthetic-based person authentication system based on deep learning architecture. In this paper, the advantages of deep learning and machine learning architectures are leveraged for automatic music aesthetic feature extraction and making predictions with reduced computational cost.

III. METHODOLOGY

A. OVERVIEW

This research proposes an audio aesthetic system that uniquely identifies users based on their preferred music set. Initially, the raw mp3 data was converted into audio waveforms. The generated audio waveforms were further converted into mel-spectrograms, a logarithmic transformation of an audio signal's frequency by utilizing mel-scales [25]. Mel-spectrograms were chosen as sounds of equal distance represented using mel-spectogram are perceived to be of equal distance to humans [26]. Later, a pre-trained convolutional neural network, Residual Network (ResNet), was used to extract high-dimensional feature vector from the mel-spectrogram. The skip connection property of the residual network amplifies the feature maps extracted from the mel-spectograms, thus extracting more discriminative features. Following that, a novel hybrid meta-heuristic approach called Cuckoo Search based Whale Optimization Algorithm (CSWOA) was proposed to select the most discriminatory and effective feature set from the high-dimensional feature vector. A combination of 3-set songs was generated from each user's chosen music set to generate unique templates for each user. Then, the obtained data samples for each user were fed into XGBoost classifiers for final prediction. The overall architecture of the proposed system is depicted in Fig. 1.

B. GENERATION OF AUDIO WAVEFORM AND MEL-SPECTROGRAM

Initially, the audio file was converted into a digital representation of an audio signal sampled at regular time intervals and by considering the amplitude at each sample. The default sampling rate for audio datasets of 22050 Hertz was used. After that, each audio waveform was trimmed to eliminate the silent portions of each audio clip. To ensure a static input dimension, all the audio clips were zero-padded to an equal length of 30 seconds. As deep learning architectures do not take raw audio waveform directly as input, each processed audio waveform was transformed into its mel-spectrogram representation.

Mel-spectrogram is a modified version of the spectrogram [26]. A spectrogram is a visual representation of the signal strength or the loudness over time at different frequencies represented in a specific waveform. The third dimension of the spectrogram represents amplitude with varying brightness. In a spectrogram, loud events appear as bright colours, and quiet events appear as dark colours as illustrated in Fig. 4. The vertical and horizontal axis of a spectrogram plot represent frequency and time, respectively. Spectrograms are generated using Fourier transformation of audio signals [26]. Most of the time, spectrograms use a linear scale to measure frequency. On the other hand, humans perceive frequency as a logarithmic scale [26]. Consequently, there is insufficient information retained from spectrograms for training deep learning models [26]. To resolve this issue, mel-spectrograms are considered in this research for extracting more discriminating features. Relative to the regular spectrograms, mel-spectrograms utilize mel-scale and decibel scale to measure frequency and amplitude, respectively. Mel scale is a logarithmic conversion of an audio signal's frequency [26]. Following is the formula of conversion between hertz and mel-scale [27]:

$$m = Clog(1 + \frac{f}{fc}) \tag{1}$$

Here, m denotes frequency in mel scale, f denotes frequency in linear scale, f_c is the corner frequency where the scale changes from linear to logarithmic, C is a constant that is chosen such that 1000 Hz to 1000 mel. Usually, the value of corner frequency, f_c is fixed at 700. If natural logarithm is considered, the value of the constant C is 1127, and in case of considering logarithm with base 10, the value of the constant C is 2595.

From Fig. 2, it can be observed that lower frequencies in hertz correspond to the higher distance between mels. On the other hand, higher frequencies in hertz have a smaller distance between mels. This property reinforces mels human-like perception.

Mel-spectrograms provide the deep learning architecture with similar information to what a human would perceive. The raw audio waveforms are passed through mel filter banks to obtain the mel-spectrogram. Mel filter banks are used to map frequency bins from Short Term Fourier Transform (STFT) to Mel bins. After this, each sample received a shape of 500 $\times$ 500, indicating 500 mel filter banks and 500 time steps per clip. Fig. 3 and Fig. 4 depict the visualization of a sample audio waveform and its corresponding mel-spectrogram.

C. FEATURE EXTRACTION AND OPTIMAL FEATURE SUBSET SELECTION

Fig. 5 depicts an overall diagram for the sub-optimal feature set selection by hybridizing Cuckoo search and Whale optimization algorithm. The extracted high-level features from the pre-trained residual network are separately passed to Cuckoo search and Whale optimization algorithm. After that, both algorithms produce the best population according to their method. The hybridization method contains three steps. The first step is combining the most optimal population by computing the significance of the features contributing most to the population sets. The Average Weighted Combination Method (AWCM) is employed for computing the importance of the feature subset. The second step of the hybridization is to compute a threshold value, AWCM cutoff, to obtain the new optimized feature vector. The third step of the hybridization

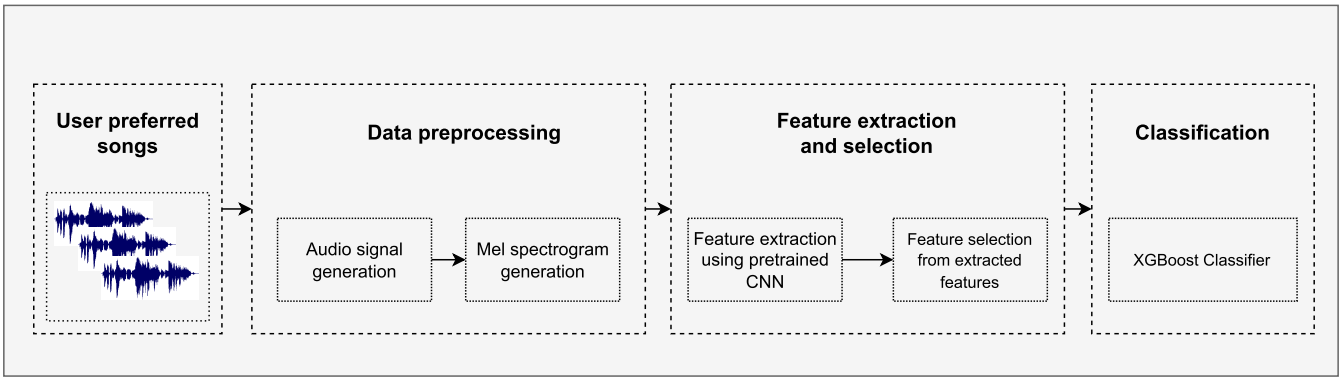

FIGURE 1. Overall flowchart of the proposed system.

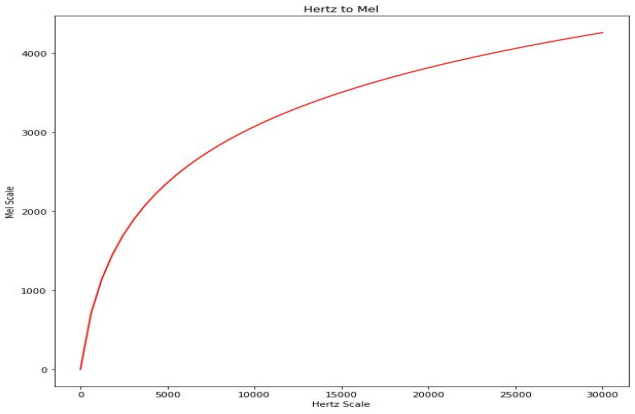

FIGURE 2. Transformation between hertz and mels.

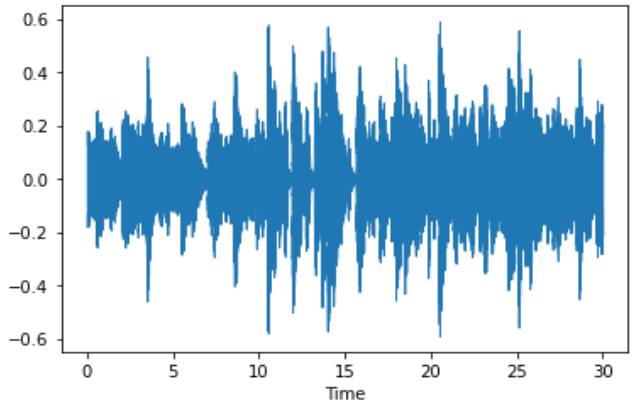

FIGURE 3. Sample of raw audio waveform generated from the mp3 data.

is to apply Sequential One-Point Flipping (SOPF) algorithm to eliminate the redundant and unnecessary features from the newly produced feature vector. After removing the redundant features from the feature set, the sub-optimal feature set is obtained.

1) DESIGN OF HIGH-LEVEL FEATURE EXTRACTION

In this phase of the proposed system, high-level features were extracted from mel-spectrograms using a CNN architecture that is pre-trained. Using traditional feature

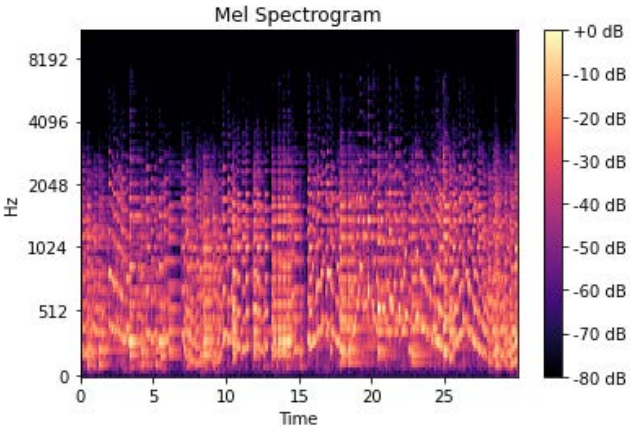

FIGURE 4. Sample of mel-spectrogram generated from the raw audio waveform.

engineering methods, generating a significant feature vector might be challenging when the associated dataset is large. Residual Network, a pre-trained CNN architecture, was proposed in this study to resolve this issue. He *et al.* [28] was the first to examine the vanishing gradient problem in an extremely deep convolutional neural network. Vanishing gradient occurs during the backpropagation phase of neural network training. During each iteration of neural network training, the weights are updated proportionally to the loss function's partial derivative with respect to the current weights. However, during the training of a very deep neural network, the gradient becomes vanishingly small, preventing the weights from changing. As a result, it may stop the neural network from further training [29]. The authors proved that if a CNN architecture comprises many layers, it will fail to generalize during the optimization process. They suggested including residual blocks or skip connections into the CNN architecture to aid this issue.

The concept of residual connection is to propagate a layer's output feature maps to its immediate layers and layers that follow. This method assists the design in aggregating data across the network, hence mitigating the problem of disappearing gradients. Residual networks have been used in

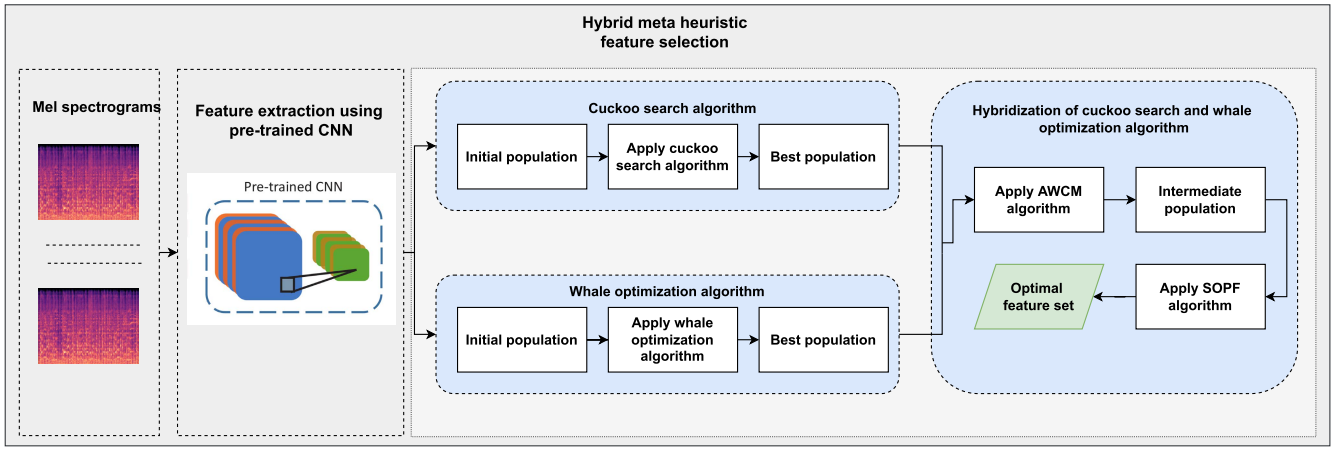

FIGURE 5. Flowchart of the proposed feature extraction and optimal feature subset selection method.

several applications, including image segmentation, medical image analysis, robotics, etc. This CNN architecture employs the pre-trained weights to extract high-level image features from dimension $224 \times 224 \times 3$. Since the mel-spectrograms created in the previous phase had dimensions of 500×500, it is essential to change the mel-spectrograms into a dimension compatible with the pre-trained CNN architecture.

For this purpose, the mel-spectrograms were downsized to 224×224 pixels. As the residual network requires RGB images as input, the single-channel mel-spectrogram images must be transformed to three-channel images. The mel-spectograms were stacked three times to rectify this, resulting in an RGB representation of the mel-spectrogram. After passing the mel spectrograms through the residual network, they generate updated feature maps, which are then flattened into a one-dimensional, fully connected neural network. This residual network comprises three fully connected layers with a size of 2048, 2048, and 1000, respectively [28]. The second last fully connected layer of the residual network was employed as the intermediate embedding feature vector of the mel-spectrograms, which is the high-level feature representation of the mel spectrograms. Therefore, the dimension of the features extracted from the residual network was 2048. Using the residual network pre-trained model serves to represent each of the chosen colour images of a mel-spectrogram through a lower representation by utilizing its high-level feature map.

2) LOW-DIMENSIONAL REPRESENTATION USING HYBRID META-HEURISTIC FEATURE SELECTION ALGORITHM

The generated high-dimensional feature vector from ResNet may induce overfitting during classification, resulting in redundant features. In order to overcome this issue, a hybrid meta-heuristic feature selection algorithm was proposed. Meta-heuristic feature selection algorithms are generic search-based optimization algorithms capable of finding the optimal feature subset from a large feature space. The purpose of these algorithms is to eliminate the inconsistent and redundant feature sets to increase classification accuracy and reduce inference time and memory consumption. The most commonly used feature selection algorithms are Genetic Algorithm (GA) [30], Particle Swarm Optimization (PSO) [31], Cuckoo Search Algorithm (CS) [32], and Whale Optimization Algorithm (WOA) [33]. In this work, a hybridization of Cuckoo Search (CS) and Whale Optimization (WO) algorithms was proposed for selecting an optimal feature subset.

a: CUCKOO SEARCH ALGORITHM

Cuckoo Search (CS) algorithm is one of the most popular meta-heuristic algorithms, inspired by the characteristic of some cuckoo species [32]. The cuckoo birds breed in the nest of other bird species called host species. This algorithm aims to increase the survival and productivity of the cuckoo birds by helping them not get discovered by the host birds. Following are the three major criteria for CS implementation [32]:

- Each cuckoo lays only one egg at once and puts it in an arbitrarily selected nest.
- The best nest with high quality eggs will carry over to the next generation
- The host nest count is constant, and host bird's probability of identifying a cuckoo's egg is $p_{hb} \in (0, 1)$. Host bird can throw the egg or abandon the nest, as well as build a new nest.

Here, the imitation of the above natural phenomena is that the eggs in a nest represent a set of solutions, while new cuckoo egg suggests a new solution. High-quality eggs represent the best optimal solution. This means that eggs that resemble the host birds can hatch and mature without being discovered by the host birds. Thus, the less fit solution will be replaced by the new and better one. The number of host nests represents the population. The fraction of cuckoo eggs discovered by the host bird is discarded, while the remaining are retained as optimal solutions for the following generation.

Another significant parameter of this algorithm is Levy Flight referred to as the random walk of cuckoo for generating a new solution (egg) in the host nest. After laying a new egg at the position of an arbitrarily picked egg, a Levy Flight is initiated. If the newly selected location is fitter than another chosen egg's location randomly, this same egg is relocated to the newly elected nest. Using the Levy Flights search algorithm, the CS algorithm may concurrently obtain all optima in a solution space. This behavior has been applied to optimization and optimum search problems, and preliminary findings indicate its elevated potential [34]. To prevent becoming trapped in the local optimum, far-field randomization introduces a significant proportion of new solutions whose locations are sufficiently distant from the existing best solution. The following is the equation for Levy Flight [32]:

$$s_i(n+1) = s_i(n) + \alpha \oplus Levy(\lambda) \tag{2}$$

In equation 2, s_i^{n+1}, s_i^n, α, $\oplus$ and $Levy(\lambda)$ represent new solutions, current location, step size,entry wise multiplication during walk and Levy exponent respectively. The Levy flight is basically a random walk, with random step length determined by a Levy distribution with infinite variance and mean. The formula of $Levy(\lambda)$ is the following [32]:

$$Levy \sim u = n^{\lambda}, (1 < \lambda \leq 3) \tag{3}$$

In equation 3, u is a normal stochastic variable and n represents current iteration. In this work, the number of features selected by the Cuckoo Search algorithm was 618.

b: WHALE OPTIMIZATION ALGORITHM

This algorithm is inspired by the bubble net feeding strategy of humpback whales searching for food [33]. WOA is a widely used meta-heuristic algorithm that has been demonstrated to be effective, straightforward to implement, and capable of generating robust and relevant feature subsets [35]. The whale behavior in this optimization algorithm is to search for prey, encircling, and attacking it. The search for prey is referred to as the exploration phase. Once the target is discovered, they begin their attack by encircling it.

Since the optimal solution in the search space is unknown at the beginning of the exploration phase, the agent (the whale) chooses a random prey as the current best option. Once an agent identifies the optimal solution, the other agents will update their locations by directing toward the optimal option. The equations 4 and 5 represent the prey localization and encircling method of whale [33]:

$$d = |c * y'(n) - y(n)| \tag{4}$$

$$y(n+1) = |y'(n) - Z * d| \tag{5}$$

Here, Z and c are the co-efficients. y' is the position vector of the best solution obtained so far. y is the current position vector at n_{th} iteration.

Z and c are updated using the following equations [33]:

$$c = 2 * r \tag{6}$$

$$Z = 2 * a * r - a \tag{7}$$

$$a = \frac{2 - 2 * n}{n_{max}} \tag{8}$$

Here, a is a convergence factor that reduces from 2 to 0 over the iterations, r is a random value ranging between [0,1], n_{max} denotes maximum number of iterations.

Shrinking encircling mechanism and spiral updating positions are the two types of prey hunting mechanisms to update the whale's position for finding an optimal solution. These methods depend on a probability factor $'p'$. In the Shrinking encircling mechanism, the optimal solution is obtained by reducing the value of a, while in the spiral updating positions, a whale travels spirally to reach the destination. In the spiral updating mechanism, to update the location of the spiral, the distance between the whale (X', Y') and the prey (X, Y) is calculated. The movement of the spiral shape is updated using the following equations [33]:

$$y(n+1) = \begin{cases} y'(n) - Z * d & if\ p < 0.5 \\ d' * e^{bu} * cos(2\pi u + y'(n)) & if\ p \geqslant 0 \end{cases} \tag{9}$$

Here, p is a random number ranging in [0,1], b is a constant that clarify the shape of the spiral and u is a random number between $[-1,1]$ and $d' = |y' - y|$.

During prey exploration, if $|Z| > 1$, then a random search agent (y_{rand}) is chosen from the entire population and the location of the current search agent is updated by equations 10 and 11. On the other hand, if $|Z| < 1$, the whales move towards the global best solution and its position is updated by equations 4 and 5. The equation of prey search is the following [33]:

$$d = |c * y_{rand}(n) - y(n))| \tag{10}$$

$$x(n+1) = |y_{rand}(n) - Z * d)| \tag{11}$$

The exploitation phase depends on the distance between a search agent and the best search agent. If some random search agents are far away from the global solution, then the convergence time of the algorithm increases slightly [36]. In this work, the number of features selected by the Whale Optimization algorithm was 495.

c: CSWOA: HYBRIDIZATION OF CS AND WOA

Cuckoo Search and Whale Optimization Algorithms are two unique meta-heuristic feature selection algorithms. Due to few tuning parameters, the CS algorithm is simple to implement and can converge promptly [37]. On the other hand, the WOA algorithm does not get trapped in local optima and thus rapidly converges to the optimal global solution. These factors enable the method to tackle a wide range of real-world problems without significant structural modifications [36]. The merits of both methods have been leveraged to generate a feature vector that is more optimal than either approach could obtain individually.

In the proposed hybridization approach, the best feature vector from both algorithms was initially obtained by implementing them separately. Then, the most optimal population was combined by evaluating the significance of all features pertaining to each of the two population sets. The features were represented in binary form (0 or 1). Evaluation of the importance of the feature subset was obtained by the Average Weighted Combination Method (AWCM) [38]. A threshold value, AWCM cutoff, was computed to get the new optimized feature vector. On the other hand, a non-greedy local search algorithm, called Sequential One-Point Flipping (SOPF) [38] eliminated the redundancy in the newly generated feature vector.

In the AWCM algorithm, a summation of the accuracy of all the feature vectors generated from each algorithm was determined. The resulting sum of each feature's accuracy can be referred to as a feature's importance. If the feature importance of a particular feature was greater than the AWCM cutoff, then that feature was taken. An example of a final optimized feature vector by utilizing feature importance using the AWCM algorithm is demonstrated in Table 1.

In addition, there was a possibility of the presence of redundant features in the optimal feature set obtained from AWCM. This may further reduce the classification accuracy. A local search algorithm, Sequential One Point Flipping (SOPF) algorithm was used to eliminate the redundant features to get rid of this problem [38]. This algorithm sequentially traverses each optimal feature set and inspects the effect of the neighboring feature sets on the features under consideration. It successively flips the state of each feature and calculates its fitness. That feature is accepted if any intermediate feature exhibits higher accuracy than a current solution. SOPF ensures an efficient and scalable feature vector by removing redundancy.

Algorithm 1 Sequential One Point Flipping Algorithm [38]

Input: Initial feature set, total number of features.
Output: Final optimized feature set.
$F_{initial}$=Initial feature set.
F_{mid} = Generated intermediate feature set from various combinations.
n = Total number of features.
F_{final} = Final optimized feature set.
$F_{mid} = F_{initial}$
for *i= 1 to n* **do**
  F_{temp} = flip value of feature interi in F_{mid}
  if $Accuracy(F_{temp}) > Accuracy(F_{mid})$ **then**
    $F_{mid} = F_{temp}$
end
$F_{final} = F_{mid}$
Output: Final feature set

The obtained size of the final feature vector generated from the hybrid CSWOA algorithm was 532. These feature vectors were considered as the most discriminating feature vectors that will be passed into machine learning classifiers for the final prediction.

D. CLASSIFICATION BLOCK

The classification block of the proposed method aims to perform identification of the users based on their audio-aesthetic preferences. After generating a feature vector for each user, testing and training sets were generated for classification. In this work, a unique 3-songs set was generated using the combination method $\binom{10}{3}$ from the total number of user-liked songs while retaining the user's aesthetic preferences. The combination method assured no duplicate sets of data points within the population and that similar data points with different song orders are discarded. After performing combination the total datapoints for FMA dataset became 4080 and for MPD dataset the value was 24000. In addition, after combination the size of each datapoints feature vector became 532×3. Transformation of the dimensionality of the original audio file for every user is depicted in Fig. 6. First, the raw audio song is transformed into an audio signal which is of dimension 22050×30. Later, the audio signal was converted into an RGB mel-spectrogram of dimension 224×224x3. This transformation is a part of pre-processing for passing the mel-spectrogram into the residual network. The pretrained residual network extracts 2048-dimensional feature vector which is further reduced to 532 by using feature selection algorithm.

For classification, XGBoost was used to identify the users with the highest precision. XGBoost stands for eXtreme Gradient Boosting [39], which utilizes a parallel tree gradient boosting mechanism. This algorithm supports three types of gradient boosting methods [39]:

- **Gradient Boosting:** Referred as gradient boosting machine that only deals with the learning rate.
- **Stochastic Gradient Boosting:** Sub-samples at the row, column, and column per split levels.
- **Regularized Gradient Boosting:** Leverages the advantages of L1 and L2 regularization.

This XGBoost classifier yields superior results by efficiently using memory resources with less computational time. It uses Sparse Aware, which can automatically handle missing values. In addition, it contains a Block Structure that facilitates parallelization during tree construction. Furthermore, it performs continuous training, which can improve a prefitted model's performance on unknown data points [39].

The XGBoost classifier boosts gradients using the Gradient Boosting Decision Tree technique. It is a type of ensemble learning that enables sequential model embedding until performance hits convergence. These characteristics make this algorithm more robust. The advantages of this algorithm over others are its fast execution time, scalable kernel, and extensive selection of adjustable hyperparameters. Moreover, in this work, it outperformed other conventional classifiers, as demonstrated by the experiments described in the following section.

TABLE 1. Example of the final feature vector optimization using AWCM algorithm.

Name of feature selection algorithms	Population	f1	f2	f3	f4	Accuracy	w1	w2	w3	w4
Cuckoo Search Algorithm (CSA)	CS1	1	1	0	1	0.82	0.82	0.82	0	0.82
	CS2	0	1	1	1	0.88	0	0.88	0.88	0.88
	CS3	1	0	1	1	0.75	0.75	0	0.75	0.75
	CS4	1	0	1	0	0.85	0.85	0	0.85	0
	CS5	1	1	1	0	0.97	0.97	0.97	0.97	0
Whale Optimization Algorithm (WOA)	WOA1	1	0	0	1	0.78	0.78	0	0	0.78
	WOA2	1	1	0	0	0.72	0.72	0.72	0	0
	WOA3	0	1	0	1	0.89	0	0.89	0	0.89
	WOA4	0	1	1	1	0.95	0	0.95	0.95	0.95
	WOA5	0	1	1	1	0.85	0	0.85	0.85	0.85
Sum of feature importance							4.89	6.08	5.25	5.92
AWCM cutoff ($(\frac{1}{n}\sum_{i=1}^{n} W_i)$)							5.405			
Final feature set							0	1	0	1

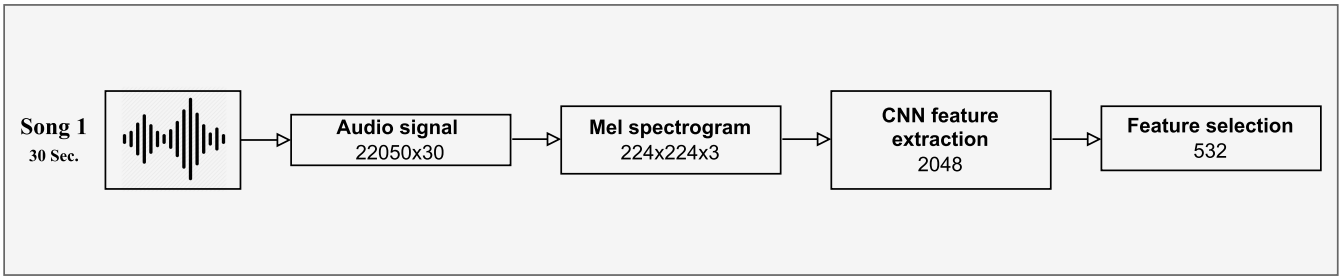

FIGURE 6. High level and low level feature dimension before concatenating them for passing through final classifier.

IV. EXPERIMENTAL RESULTS

A. DATASET DESCRIPTION

The performance of the proposed three-stage audio aesthetic framework was evaluated on two datasets; one is proprietory [5], and another is publicly available [40]. The proprietary dataset consists of 34 users and their corresponding ten favorite songs. The users chose the songs from a set of 224 songs. The 224 songs were collected from Free Music Archive (FMA). The original FMA dataset consists of 917 gigabytes of Creative Commons-licensed tracks, and 161 genres [41]. The collected songs had a balanced mix of different music genres such as Pop, Rock, Folk, Hip-Hop, Jazz, Country, Classical, and Disco. Another dataset was constructed using publicly available the Million Playlist Dataset (MPD) from Spotify. There are 200 anonymous Spotify users' playlists sampled from the first 1000 MPD playlists. Each playlist consists of 10 songs, each with a 30-second song clip. The dataset was divided into 70:30 for training and testing.

To evaluate the model's performance, different evaluation parameters such as accuracy, precision, recall, and F1-score were utilized [42]. These evaluation metrics depend on the parameters of the confusion matrix. A confusion matrix measures the performance of machine learning classifiers visualizing the actual and predicted results by a classifier [42]. The parameters associated with it are,

- **True Positive** (t_p)**:** The percentage of positive predictions, that were actually positive.
- **True Negative** (t_n)**:** The percentage of negative predictions, that were actually negative.
- **False Positive** (f_p)**:** The percentage of positive predictions, that were actually negative.
- **False Negative** (f_n)**:** The percentage of negative predictions, that were actually positive.

The above parameters are used generally for binary classifications, but they can be derived for multi-class classifications as well. The equations for the evaluation metrics are given below [42]:

$$accuracy = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \tag{12}$$

$$precision = \frac{t_p}{t_p + f_p} \tag{13}$$

$$recall = \frac{t_p}{t_p + f_n} \tag{14}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{15}$$

B. EXPERIMENTAL SETUP

The proposed method was implemented using Python programming language. For the machine learning algorithms sci-kit learn library was employed. The machine learning models were trained on a Corei5 CPU and NVIDIA 1080 GTX GPU backend. For the residual network the default architecture was used. For the cuckoo search algorithm, the number of nests was set to 20, step length was set to

TABLE 2. Performance analysis of ResNet with VGG16 and InceptionNet.

Dataset	Architecture	Accuracy (%)
FMA	InceptionNet + CSWOA +XGBoost	97.28
	VGG16 + CSWOA +XGBoost	96.03
	ResNet + CSWOA +XGBoost	**99.54**
MPD	InceptionNet + CSWOA +XGBoost	97.45
	VGG16 + CSWOA + XGBoost	97.38
	ResNet + CSWOA + XGBoost	**99.79**

0.01, and the levy distribution parameter was set to 1.5 [43]. For the whale optimization algorithm, the population size was set to 100 with random search ability of 0.1 [44]. For the XGBoost classifier alpha was set to 0.2, max depth of the tree was set to 5, and number of estimators for boosting was set to 1000 [45]. These parameters ensure algorithms' optimal performance.

C. EFFICACY OF DIFFERENT COMPONENTS OF THE PROPOSED METHOD

The first set of experiments demonstrated the efficiency of the different components of the proposed method. The first experiment was conducted to show the strength of the residual network that was employed as the feature extractor of the proposed method. The second experiment established the importance of the hybrid meta-heuristic feature selection algorithm. The final experiment was performed to show the efficiency of the XGBoost classifier for final classification.

1) EFFICACY OF THE RESIDUAL NETWORK

The first step of the proposed architecture was to extract high-level features using a pre-trained CNN architecture. Several different pre-trained architectures were considered. Among them, ResNet [28], VGG-16 [46], and InceptionNet [47] were selected as the most used pre-trained architectures. Table 2 illustrates the performance of the architecture with ResNet. The proposed method with the ResNet as the feature extractor attained 2.25% higher accuracy than InceptionNet and 3.51% higher accuracy than VGG-16 on the FMA dataset. On the other hand, for the MPD dataset, 2.34% higher accuracy than InceptionNet and 2.41% higher accuracy than VGG-16 were achieved. There are several reasons for the superiority of ResNet performance over others. The key advantage is addressing the problem of diminishing gradients via the skip connections method. This skip connection bypasses a few layers during training and connects directly to the output. Regularization allows any layer to be skipped if it degrades the architecture's performance. Therefore, in ResNet, a very deep neural network is trained without the vanishing gradient issue.

2) EFFICACY OF THE PROPOSED HYBRID META-HEURISTIC FEATURE SELECTION ALGORITHM

In this work, a hybrid meta-heuristic feature selection algorithm (CSWOA) is proposed to select the optimal feature subset from the entire feature set generated from pre-trained ResNet. To show the efficiency of the proposed feature selection method, an extensive comparison was performed with some standalone feature selection algorithms. The proposed hybrid meta-heuristic feature selection algorithm was compared with Particle Swarm Optimization (PSO) [31], Equilibrium Optimization (EO) [48], Cuckoo Search (CS) [32], and Whale Optimization (WO) [33]. In addition, hybridization of Particle Swarm Optimization and Equilibrium Optimization (PSEO) algorithms was also implemented for comparison. All the above-mentioned feature selection algorithms are wrapper-based methods. From Table 3, it is observed that the proposed method with CSWOA feature selection algorithm attained 2.47% higher accuracy than PSO, 3.33% higher accuracy than EO, 0.92% higher accuracy than CS, 1.09% higher accuracy than WOA, and 1.65% higher accuracy than PSEO algorithm on FMA dataset. On the other hand, for the MPD dataset, the proposed CSWOA algorithm achieved 1.51% higher accuracy than PSO, 1.54% higher accuracy than EO, 1.31% higher accuracy than CS, 1.09% higher accuracy than WO, and 1.17% higher accuracy than PSEO algorithm. The CS algorithm uses very few parameters, leading to its easy implementation and fast convergence. On the other hand, the WOA algorithm can rapidly find the global optimal solution, as it does not get stuck in local optima. The proposed hybrid meta-heuristic feature selection algorithm combines the most optimal feature subsets of two different wrapper-based algorithms and produces the optimal subset by using AWCM and SOPF methods. Therefore, the proposed hybrid meta-heuristic feature selection algorithm has higher accuracy than standalone feature selection algorithms. Fig. 7 and Fig. 8 demonstrate the results of different feature selection algorithms on FMA and MPD datasets, respectively.

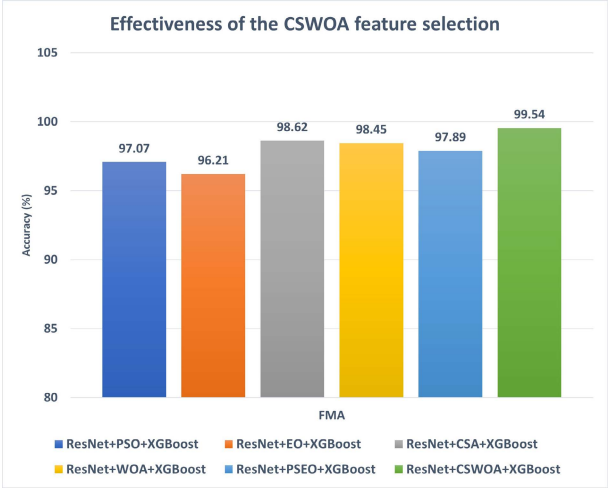

FIGURE 7. Comparison of CSWOA feature selection algorithm with PSO, EO, CSA, WOA,PSEO on FMA dataset.

3) EFFICACY OF THE XGBoost CLASSIFIER

For the final classification, XGBoost classifier was employed. In this experiment, XGBoost classifier was compared with Naive Bayes [49], Random Forest [50], Support

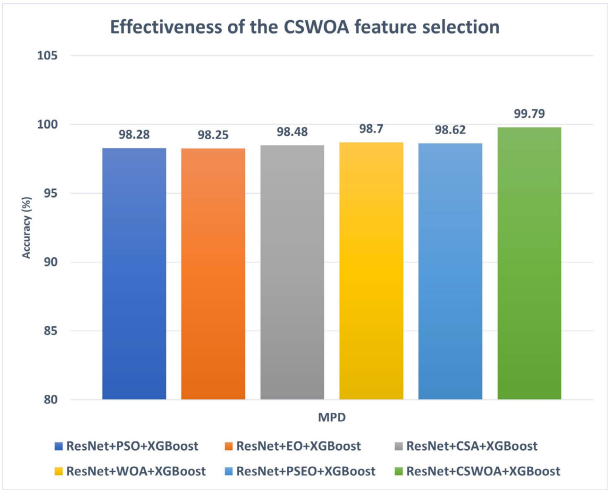

FIGURE 8. Comparison of CSWOA feature selection algorithm with PSO, EO, CSA, WOA,PSEO on MPD dataset.

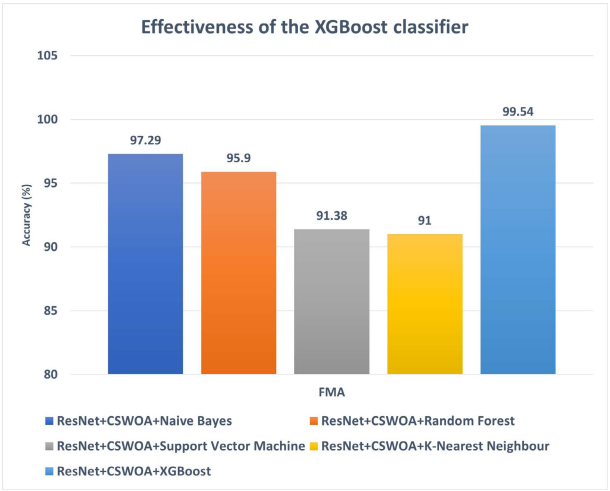

FIGURE 9. Comparison of XGBoost classifier with Naive Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbour on FMA dataset.

TABLE 3. Performance analysis of CSWOA hybrid meta-heuristic feature selection algorithm with PSO, EO, CSA, WOA, and PSEO algorithms.

DataSet	Architecture	Accuracy(%)
FMA	ResNet+PSO+XGBoost	97.07
	ResNet+EO+XGBoost	96.21
	ResNet+CSA+XGBoost	98.62
	ResNet+WOA+XGBoost	98.45
	ResNet+PSEO+XGBoost	97.89
	ResNet+CSWOA+XGBoost	**99.54**
MPD	ResNet+PSO+XGBoost	98.28
	ResNet+EO+XGBoost	98.25
	ResNet+CSA+XGBoost	98.48
	ResNet+WOA+XGBoost	98.70
	ResNet+PSEO+XGBoost	98.62
	ResNet+CSWOA+XGBoost	**99.79**

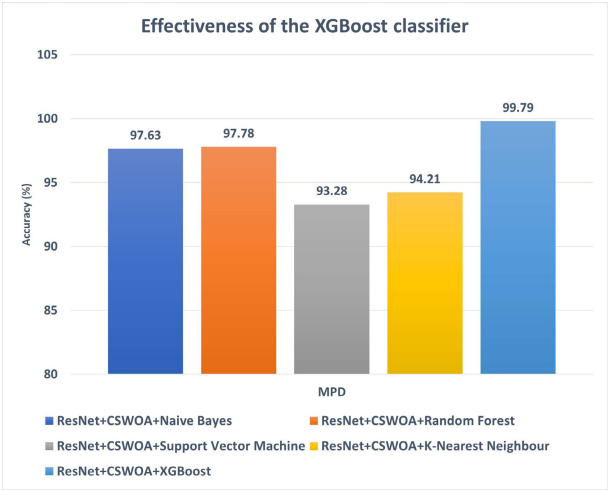

FIGURE 10. Comparison of XGBoost classifier with Naive Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbour on MPD dataset.

TABLE 4. Performance analysis of XGBoost classifier with Naive Bayes, Random Forest, Support Vector Machine and K-Nearest Neighbour.

Dataset	Architecture	Accuracy (%)
FMA	ResNet+CSWOA+Naive Bayes	97.29
	ResNet+CSWOA+Random Forest	95.9
	ResNet+CSWOA+Support Vector Machine	91.38
	ResNet+CSWOA+K-Nearest Neighbour	91
	ResNet+CSWOA+XGBoost Classifier	**99.54**
MPD	ResNet+CSWOA+Naive Bayes	97.63
	ResNet+CSWOA+Random Forest	97.78
	ResNet+CSWOA+Support Vector Machine	93.28
	ResNet+CSWOA+K-Nearest Neighbour	94.21
	ResNet+CSWOA+XGBoost Classifier	**99.79**

Vector Machine [51], and K-Nearest Neighbour [52]. Table 4 demonstrates the superiority of XGBoost over other classifier. XGBoost attained 2.25% higher accuracy than Naive Bayes, 3.64% higher accuracy than Random Forest, 8.16% higher accuracy than Support Vector Machine, and 8.54% higher accuracy than K-Nearest Neighbour on FMA dataset. For the MPD dataset XGBoost classifier attained 2.16% higher accuracy than Naive Bayes, 2.01% higher accuracy than Random Forest, 6.51% higher accuracy than Support Vector Machine, and 5.58% higher accuracy than K-Nearest Neighbour classifier. The XGBoost classifier utilizes the Gradient Boosting Decision Tree approach to enhance gradients. It is a form of ensemble learning that allows successive model embedding until the result converges. Because of this advantage, the XGBoost classifier attained higher accuracy than the other algorithms. Fig. 9 and Fig. 10 illustrate the results obtained from different classifiers on FMA and MPD datasets, respectively.

D. ABLATION STUDY

For the ablation study, three experiments were conducted. The first experiment was performed with a simple neural network architecture without any hybrid meta-heuristic feature selection algorithm and machine learning classifier. The second experiment was performed without any hybrid

TABLE 5. Ablation study for different components of the proposed method on FMA dataset.

Dataset	Experiment No.	Architecture		Accuracy (%)
FMA	1	ResNet	✓	96.23
		CSWOA	×	
		XGBoost Classifier	×	
	2	ResNet	✓	90.45
		CSWOA	×	
		XGBoost Classifier	✓	
	3	ResNet	✓	97.38
		CSWOA	✓	
		XGBoost Classifier	×	
	4	ResNet	✓	99.54
		CSWOA	✓	
		XGBoost Classifier	✓	

TABLE 6. Ablation study for different components of the proposed method on MPD dataset.

Dataset	Experiment No.	Architecture		Accuracy (%)
MPD	1	ResNet	✓	96.88
		CSWOA	×	
		XGBoost Classifier	×	
	2	ResNet	✓	91.89
		CSWOA	×	
		XGBoost Classifier	✓	
	3	ResNet	✓	98.29
		CSWOA	✓	
		XGBoost Classifier	×	
	4	ResNet	✓	99.79
		CSWOA	✓	
		XGBoost Classifier	✓	

meta-heuristic feature selection algorithm. In this experiment, a pre-trained residual network was used to extract features. Later, those features were passed into the XGBoost classifier. The final experiment was conducted by removing the XGBoost classifier from the proposed method but retaining the residual network and hybrid meta-heuristic feature selection algorithm. The final classification was performed using a simple Multi-Layer Perceptron.

1) PROPOSED METHOD WITHOUT HYBRID META-HEURISTIC FEATURE SELECTION AND XGBoost CLASSIFIER

The first ablation experiment on both datasets was the implementation of ResNet for identification. In the first row of Table 5 and Table 6, the illustration of this configuration is observed. This approach obtained 96.23% and 96.88% accuracy on FMA and MPD datasets, respectively. Compared to the proposed architecture, the accuracy of this method drops by 3.31% on the FMA dataset and 2.91% on MPD datasets. The reason for the performance drop with only residual connection is that in this experiment no feature selection method was used.

2) PROPOSED METHOD WITHOUT HYBRID META-HEURISTIC FEATURE SELECTION

The second ablation study eliminated the hybrid feature selection algorithm while retaining ResNet for feature extraction and XGBoost classifier for classification. From the second row of Table 5 and Table 6, it is observed that the accuracy of this approach was reduced by 9.06% and 7.9% for FMA and MPD datasets, respectively. In this experiment, the high-level features were extracted by a pretrained convolutional neural network and classification was done by XGBoost classifier. The high-level features extracted from the pretrained CNN contain complex feature representation, which a classical machine learning algorithm is unable to handle accurately. Therefore, for this experiment the performance was dropped by 9.09% for the FMA dataset and 7.9% for the MPD dataset.

3) PROPOSED METHOD WITHOUT XGBoost CLASSIFIER

In the third ablation study, experiment was conducted without XGBoost classifier. This experiment extracted features using ResNet, selected the optimal feature set using CSWOA, and used MLP instead of XGBoost for classification. This experiment also illustrates the dominance of the proposed architecture. The accuracy of this approach is reduced by 2.16% for FMA dataset and 1.5% for MPD dataset, when compared to the original results. This configuration is shown in the third row of Table 5 and Table 6. In the third experiment, the pretrained CNN was used to extract high level features and later a Multi-Layer Perceptron (MLP) was used for classification. The MLP achieves higher accuracy than using a classical machine learning algorithm, but the performance is still lower by 2.18% for FMA dataset and lower by 1.5% for the MPD dataset vs the proposed architecture.

E. PERFORMANCE COMPARISON WITH STATE-OF-THE ART RESULTS

Table 7 illustrates a comparative study of the previously developed audio aesthetic model with the proposed method. This comparison reveals that the proposed method has outperformed the previous method in terms of accuracy and inference time. The proposed system is able to achieve state-of-the-art results on both FMA and MPD datasets by attaining 99.54% and 99.79% rank 1 accuracy, respectively. Regarding inference time, for FMA dataset, it is 1.67s, while for MPD dataset the inference value is 7.59s. The songs in the FMA dataset are collected from a pool of 224 songs, but songs in the MPD dataset are not restricted. Thus, the improvement in accuracy for the MPD dataset compared to the FMA dataset is due to a greater song diversity, resulting in more distinctive extracted features. The user count in FMA dataset is 34, while for MPD dataset it is 200.

The audio aesthetic system introduced by Sieu and Gavrilova [5] achieved a rank 1 accuracy of 95.74% on the FMA dataset and 99.70% accuracy on the MPD dataset. Furthermore, in their work, the inference time for FMA and MPD datasets was 1.85s and 8.12s, respectively. From the above results, it is observed that the proposed approach obtained higher accuracy than the previous audio aesthetic system.

In terms of inference time, the proposed method also attained superior results by obtaining lower inference time than the previous work. The higher accuracy and lower inference time indicate that the proposed method is able to identify

users with the highest accuracy and infer unseen data faster than the previous method. Sieu and Gavrilova [5] adopted extensive feature engineering, which increased accuracy and inference time. On the other hand, in the proposed method, a pre-trained transfer learning algorithm, ResNet, was utilized for feature extraction and a hybrid meta-heuristic feature selection algorithm for optimized feature selection. Finally, classification was performed with the XGBoost classifier. The advantage of the proposed method is the implementation of ResNet and CSWOA for feature extraction and feature selection, respectively. ResNet automatically generated the most significant features, while the feature selection algorithm eliminated the redundant features and generated the most contributing feature subset. Finally, the XGBoost classifier identifies users with the highest accuracy and lowest inference time. The feature extraction and selection approach increased the accuracy by generating an optimal feature subset. In addition, XGBoost classifier helped in reducing the inference time, since it was fed to a lower number of optimal feature subset for training. From these aforementioned discussions, it can be stated that the proposed method is computationally inexpensive and provides a higher identification accuracy than the state-of-the-art method. Thus, for real-world application, the proposed approach is feasible to deploy.

TABLE 7. Comparison of Rank 1 accuracy and inference time with the previous state-of-the-art method.

System	Dataset	Accuracy (Rank 1)	Inference Time
Sieu and Gavrilova [5]	FMA	95.74%	1.85s
Proposed Method		**99.54%**	**1.67s**
Sieu and Gavrilova [5]	MPD	99.60%	8.12s
Proposed Method		**99.79%**	**7.59s**

F. RESULTS AND DISCUSSION

To establish superiority of the proposed model, several experiments were performed on both datasets. The optimized feature vector obtained from the hybrid feature selection model was fed into different machine learning classifiers for identification. XGBoost attained the highest identification accuracy for both datasets among all the classifiers. The detailed results for FMA and MPD datasets are tabulated in Table 8. From the table, it is observed that for FMA dataset, the accuracy, precision, recall, and F1-score values are 99.54%, 97%, 98%, and 99%, respectively. On the other hand, for the MPD dataset. these values are 99.79%, 98%, 99%, and 99%, respectively. The above results exhibit that the XGBoost classifier identifies users with significant accuracy and a great value of precision, recall, and F1-score.

Fig. 11 depicts the Cumulative Matching Characteristic (CMC) curve, which demonstrates the system's rank 1 to rank 5 recognition rates. The CMC curve in a person identification system reflects the system's accuracy in identifying

TABLE 8. Performance analysis of the proposed method using XGBoost classifier.

Dataset	Accuracy	Precision	Recall	F1- score
FMA	99.54%	97%	98%	99%
MPD	99.79%	98%	99%	99%

users within a specified number of predictions. A rank 1 identification rate is the value when a system correctly identifies a user in a single prediction. On the other hand, in the rank 5 identification rate, the prediction is generated within the top ten results. A CMC curve's normalized Area-Under-Curve (nAUC) measures its overall accuracy, with an ideal nAUC of 1 corresponds to a thoroughly reliable system. The proposed system obtains an nAUC of 0.9994 across all 34 user classes, with 99.54% rank 1 recognition and 100% rank 5 recognition for the FMA dataset. In contrast, the attained nAUC value for the MPD dataset is 0.9998 across all 200 users, with a rank 1 recognition rate of 99.79% and a rank 5 accuracy rate of 100%.

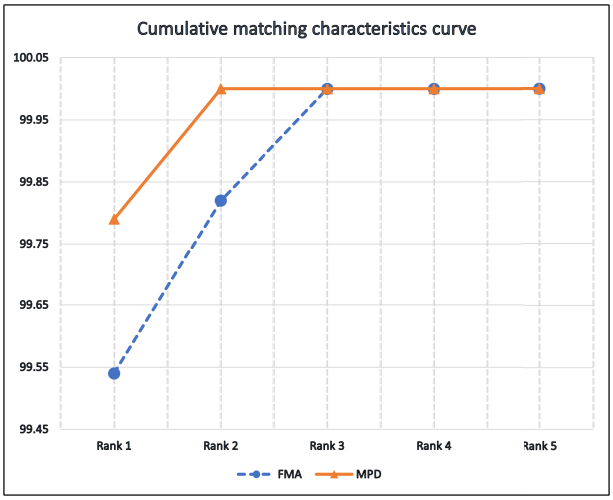

FIGURE 11. Cumulative matching characteristics curve for Rank 1 to Rank 5 accuracy for FMA and MPD dataset.

The Receiver Operating Characteristics (ROC) curve of the proposed model is depicted in Fig. 12 and Fig. 13 for FMA and MPD datasets, respectively. ROC curve is a measure of the performance of classifiers at the different thresholds. It is a plot between True Positive Rate (TPR) and False Positive Rate (FPR). A system with a high TPR followed by a low FPR has lower verification errors, indicating its reliability. In ROC curve, the accuracy of a system is measured by Area Under the Curve (AUC) score. A model's AUC value near 1 means the model is highly capable of distinguishing between different classes. Fig. 12 exhibits the AUC value of the system on FMA dataset, which is 0.995, and Fig. 13 depicts the AUC value of the system for the MPD dataset, which is 0.998. The AUC values exhibit that the proposed system is highly capable of identifying users with a significant accuracy.

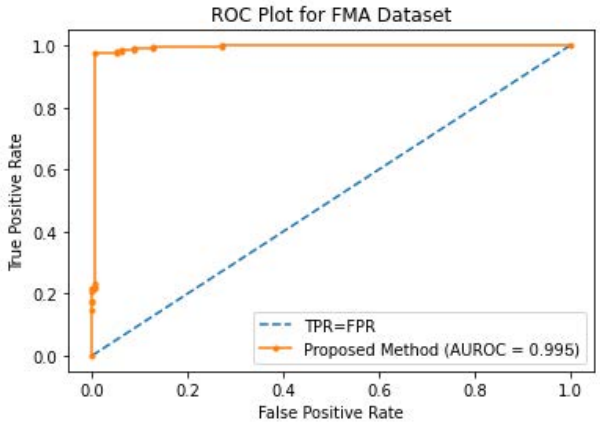

FIGURE 12. The Receiver Operating Characteristic (ROC) curve for FMA dataset.

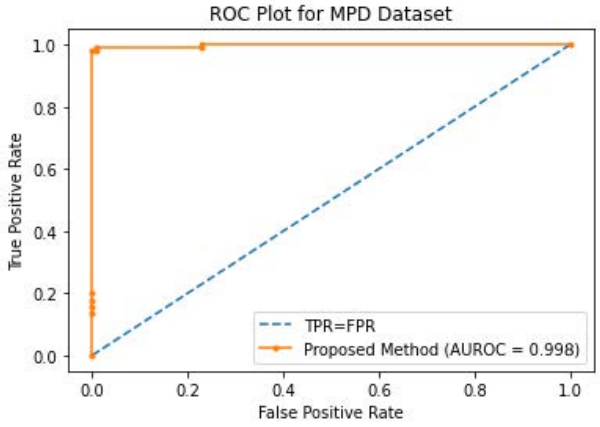

FIGURE 13. The Receiver Operating Characteristic (ROC) curve for MPD dataset.

V. CONCLUSION AND FUTURE WORKS

In this work, a novel three-stage audio aesthetic-based biometric system is proposed. To the best of our knowledge, this is the first audio aesthetic system that utilizes automatically extracted features using a deep learning approach. In the first stage of this system, deep features were extracted using pre-trained ResNet architecture. In the second stage, an optimized feature subset was chosen from the generated high-level feature set by utilizing a hybrid feature selection algorithm. In the final step, XGBoost classifier was used for user identification based on personal audio preferences. The proposed approach has achieved a rank 1 accuracy of 99.54% and 99.79% accuracy on FMA and MPD datasets, surpassing other methods. Furthermore, the nAUC scores ranging from rank 1 to rank 5 on the CMC curve further validate the system's reliability and efficacy. Experimental results demonstrate that the proposed framework can extract more discriminating features from a set of user-preferred songs. This three-stage framework proves the efficacy of intermediary features automatically extracted from deep learning architecture without extensive feature engineering.

In the future, additional system performance optimization maybe achieved through substantial parallelism. Assessing the impact of various fusion techniques on the overall performance of an audio-visual system is another potential research direction. In addition, analyzing the relationship between an individual's preferred music genre and an image category might provide a new direction of research into the aesthetic-based biometric domain.

REFERENCES

[1] A. K. Jain and S. Prabhakar, "Biometric authentication," *Scholarpedia*, vol. 3, no. 6, p. 3716, 2008.
[2] M. Gavrilova, F. Ahmed, S. Azam, P. P. Paul, W. Rahman, M. Sultana, and F. T. Zohra, "Emerging trends in security system design using the concept of social behavioural biometrics," *Information Fusion for Cyber-Security Analytics* (Studies in Computational Intelligence). Cham, Switzerland: Springer, 2017, pp. 229–251.
[3] P. Lovato, M. Bicego, C. Segalin, A. Perina, N. Sebe, and M. Cristani, "Faved! biometrics: Tell me which image you like and I'll tell you who you are," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 3, pp. 364–374, Mar. 2014.
[4] 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.
[5] B. Sieu and M. L. Gavrilova, "Person identification from audio aesthetic," *IEEE Access*, vol. 9, pp. 102225–102235, 2021.
[6] 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.
[7] S. Azam and M. Gavrilova, "Soft biometric: Give me your favorite images and i will tell your gender," in *Proc. IEEE 15th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI CC)*, Aug. 2016, pp. 535–541.
[8] N. Rodriguez-Fernandez, S. Alvarez-Gonzalez, I. Santos, A. Torrente-Patiño, A. Carballal, and J. Romero, "Validation of an aesthetic assessment system for commercial tasks," *Entropy*, vol. 24, no. 1, p. 103, Jan. 2022.
[9] M. Z. Anwar, Z. Kaleem, and A. Jamalipour, "Machine learning inspired sound-based amateur drone detection for public safety applications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2526–2534, Mar. 2019.
[10] M. Boukabous and M. Azizi, "Multimodal sentiment analysis using audio and text for crime detection," in *Proc. 2nd Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET)*, Mar. 2022, pp. 1–5.
[11] 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.
[12] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.
[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[14] R. Sousa Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, and B. Maia, "'twazn me!!!; ('automatic authorship analysis of micro-blogging messages)," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2011, pp. 161–168.
[15] P. Lovato, A. Perina, N. Sebe, O. Zandoná, 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, pp. 45–56.
[16] A. S. Bari and M. L. Gavrilova, "Artificial neural network based gait recognition using Kinect sensor," *IEEE Access*, vol. 7, pp. 162708–162722, 2019.
[17] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–9, Dec. 2021.
[18] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–34, May 2018.
[19] M. De Witte, G.-J. Stams, X. Moonen, A. E. R. Bos, and S. Van Hooren, "Music therapy for stress reduction: A systematic review and meta-analysis," *Health Psychol. Rev.*, vol. 16, no. 1, pp. 134–159, Nov. 2022.
[20] A. Lamont and D. Hargreaves, "Musical preferences," in *Routledge International Handbook of Music Psychology in Education and the Community*. Evanston, IL, USA: Routledge, 2021, pp. 131–145.

[21] V. Moscato, A. Picariello, and G. Sperli, "An emotional recommender system for music," *IEEE Intell. Syst.*, vol. 36, no. 5, pp. 57–68, Sep. 2021.
[22] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, May 2018.
[23] M. J. Patel and M. I. Husain, "An approach to developing EEG-based person authentication system," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 2619–2625.
[24] J. Sooriyaarachchi, S. Seneviratne, K. Thilakarathna, and A. Y. Zomaya, "MusicID: A brainwave-based user authentication system for Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8304–8313, May 2021.
[25] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937.
[26] F. Liu, T. Shen, Z. Luo, D. Zhao, and S. Guo, "Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation," *Appl. Acoust.*, vol. 178, Jul. 2021, Art. no. 107989.
[27] D. O'Shaughnessy, *Speech Communications: Human and Machine (IEEE)*. Hyderabad, Telangana: Universities Press, 1987.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] S. Basodi, C. Ji, H. Zhang, and Y. Pan, "Gradient amplification: An efficient way to train deep neural networks," *Big Data Mining Anal.*, vol. 3, no. 3, pp. 196–207, Sep. 2020.
[30] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1998.
[31] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw. (ICNN)*, vol. 4. Nov./Dec. 1995, pp. 1942–1948.
[32] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proc. World Congr. Nature Biol. Inspired Comput. (NaBIC)*, Dec. 2009, pp. 210–214.
[33] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, May 2016.
[34] A. H. Gandomi, X.-S. Yang, and A. H. Alavi, "Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems," *Eng. Comput.*, vol. 29, no. 1, pp. 17–35, Jan. 2013.
[35] H. M. Mohammed, S. U. Umar, and T. A. Rashid, "A systematic and meta-analysis survey of whale optimization algorithm," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–25, Apr. 2019.
[36] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, Oct. 2017.
[37] L. Liu, X. Liu, N. Wang, and P. Zou, "Modified cuckoo search algorithm with variational parameters and logistic map," *Algorithms*, vol. 11, no. 3, p. 30, Mar. 2018.
[38] M. Ghosh, R. Guha, I. Alam, P. Lohariwal, D. Jalan, and R. Sarkar, "Binary genetic swarm optimization: A combination of GA and PSO for feature selection," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1598–1610, Sep. 2020.
[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[40] 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.
[41] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," 2016, *arXiv:1612.01840*.
[42] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, nos. 3–4, pp. 429–450, Dec. 2017.
[43] J. García, B. Crawford, R. Soto, C. Castro, and F. Paredes, "A K-means binarization framework applied to multidimensional knapsack problem," *Int. J. Speech Technol.*, vol. 48, no. 2, pp. 357–380, Feb. 2018.
[44] T. Vaiyapuri and H. Alaskar, "Whale optimization for wavelet-based unsupervised medical image segmentation: Application to CT and MR images," *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, p. 941, 2020.
[45] K. Loggenberg, A. Strever, B. Greyling, and N. Poona, "Modelling water stress in a Shiraz vineyard using hyperspectral imaging and machine learning," *Remote Sens.*, vol. 10, no. 2, p. 202, Jan. 2018.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[48] A. Faramarzi, M. Heidarinejad, B. Stephens, and S. Mirjalili, "Equilibrium optimizer: A novel optimization algorithm," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105190.
[49] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, vol. 3, no. 22, 2001, pp. 41–46.
[50] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1. Aug. 1995, pp. 278–282.
[51] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[52] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statistician*, vol. 46, pp. 175–185, Aug. 1992.

FARIHA IFFATH (Member, IEEE) received the B.Sc. degree in computer science and engineering from the Chittagong University of Engineering and Technology, Bangladesh, in 2019. She is currently pursuing the M.Sc. degree in computer science with the University of Calgary, Canada, under the supervision of Prof. Marina L. Gavrilova. She worked as a Lecturer with BGC Trust University, Bangladesh, from February 2020 to May 2021. She published research in the *Computers* journal (MDPI) 2021, International Conference on Intelligent Computing & Optimization 2021, Proceedings of the International Conference on Big Data, IoT, and Machine Learning 2022, and New Approaches for Multidimensional Signal Processing 2022. Her research interests include computer vision, deep learning, biometrics, and audio/visual aesthetics.

MARINA L. GAVRILOVA (Senior Member, IEEE) is currently a Full Professor at the University of Calgary and an international expert in the area of biometric security, machine learning, pattern recognition, and information fusion. She directs the Biometric Technologies Laboratory, and published over 300 books, conference proceedings, and peer-reviewed articles. Her professional excellence was recognized by the Canada Foundation for Innovation, the Killam Foundation, U Make a Difference Award, and the Order of the University of Calgary. She is the Founding Editor-in-Chief of *Transactions on Computational Sciences* (Springer), and serves on the editorial boards for the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SCIENCES, IEEE ACCESS, *The Visual Computer*, *Sensors*, and the *International Journal of Biometrics*.