electronics

MDPI

Article

A Comparative Analysis of Modeling and Predicting Perceived and Induced Emotions in Sonification

Faranak Abri [1,*], Luis Felipe Gutiérrez [1], Prerit Datta [1], David R. W. Sears [2], Akbar Siami Namin [1] and Keith S. Jones [3]

[1] Computer Science Department, Texas Tech University, Lubbock, TX 79409, USA; luis.gutierrez-espinoza@ttu.edu (L.F.G.); prerit.datta@ttu.edu (P.D.); akbar.namin@ttu.edu (A.S.N.)
[2] Department of Music, Texas Tech University, Lubbock, TX 79409, USA; david.sears@ttu.edu
[3] Department of Psychology, Texas Tech University, Lubbock, TX 79409, USA; keith.s.jones@ttu.edu
* Correspondence: faranak.abri@ttu.edu

Abstract: Sonification is the utilization of sounds to convey information about data or events. There are two types of emotions associated with sounds: (1) "perceived" emotions, in which listeners recognize the emotions expressed by the sound, and (2) "induced" emotions, in which listeners feel emotions induced by the sound. Although listeners may widely agree on the perceived emotion for a given sound, they often do not agree about the induced emotion of a given sound, so it is difficult to model induced emotions. This paper describes the development of several machine and deep learning models that predict the perceived and induced emotions associated with certain sounds, and it analyzes and compares the accuracy of those predictions. The results revealed that models built for predicting perceived emotions are more accurate than ones built for predicting induced emotions. However, the gap in predictive power between such models can be narrowed substantially through the optimization of the machine and deep learning models. This research has several applications in automated configurations of hardware devices and their integration with software components in the context of the Internet of Things, for which security is of utmost importance.

Keywords: sonification; security alarm; acoustic features; sound analysis; Internet of Things; emotion prediction; IADSE; EmoSoundscape

Citation: Abri, F.; Gutiérrez, L.F.; Datta, P.; Sears D.R.W.; Siami Namin, A.; Jones, K.S. A Comparative Analysis of Modeling and Predicting Perceived and Induced Emotions in Sonification. *Electronics* **2021**, *10*, 2519. https://doi.org/10.3390/electronics10202519

Academic Editor: Pal Varga

Received: 31 August 2021
Accepted: 6 October 2021
Published: 15 October 2021

1. Introduction

The Internet of Things (IoT) has enabled a rich landscape of interconnected ubiquitous devices capable of offering a variety of services and applications. IoT supports a gamut of sensors that are capable of recording and transmitting data from a wide variety of sources. To ensure the reliability of these interconnected devices, when inter-operating together, extensive monitoring and alarming systems are needed. Techniques and approaches such as textual warning messages, visualization (e.g., DataDog [1]), and alarming through sounds are the mainstream channels employed for communication purposes in different hardware/software platforms that include IoT. For instance, Flight Guardian [2], a flight deck warning system designed for older airplanes lacking digital warning systems, can improve flight safety by monitoring a pilot's situational awareness using real-time video analysis and underlying knowledge to generate timely speech warnings. While the use of textual data and visualizations has been explored in typical cyber–physical systems (CPSs) and the IoT, the use of sounds in these contexts is accompanied with additional complexity and may require some other analysis and comprehension before becoming the main avenue for communication. An example of such complexity is whether certain types of sounds induce certain types of emotions within the system operator. The answer to this question is important in ensuring the effectiveness of communication in such complex systems.

1.1. Sonification Applications in the IoT

The literature regarding the use of sonification in the IoT describes the many facets and versatility of sonification across several application domains. One of the primary uses of sonification in the IoT is in medical applications. IoT sensors can continuously record and monitor data from different parts of the body. Researchers have proposed the use of the IoT for remotely monitoring elderly patients' health [3]. Measurements such as heart rate, blood pressure, and body temperature [4] can be collected remotely. In the event of an accident, such as a fall, such a system can quickly alert doctors, the patient's caretakers, or both, and can sound an alarm that could alert anyone in the patient's general vicinity [5]. Researchers have also proposed a sonification system for asthma patients that can inform a patient's emergency contacts when a sudden asthma attack occurs and can activate a buzzer to alert nearby people who may be able to help [6].

Sonification has also been used as an alternate modality to learn about bodily movements and functions. Danna and Velay [7] proposed the use of sonifications of hand movements made while writing to help researchers understand the motor control needed to perform that task, which may help patients with disabilities. Likewise, Turchet [8] proposed the use of interactive sonifications to help during the therapy of patients with limited bodily movement and control. Shoes enabled with IoT sensors can collect data and be monitored remotely to give patients feedback on their gait and body movements. The authors argued that the use of sonification in therapy can help patients with motor disabilities to walk better. Researchers have also suggested sonification of electroencephalogram (EEG) data as part of brain–computer interfaces and to help understand the brain's response to auditory stimuli as a supplement to brain imaging [9].

Sonification can also be useful in promoting overall wellbeing in IoT systems through music. Quasim et al. [10] proposed an emotion-based music recommendation and classification framework (EMRCF) to recommend songs to individuals based on their mood and previous listening history. The authors proposed the analysis of facial features to predict the person's mood, from which the system would recommend songs that were pre-sorted into one of six categories, such as joyful, inspired, enthusiastic, emotional, silent, and depressed.

Timoney et al. [11] presented a summary of research in the area of IoT and music known as the Internet of Musical Things (IoMusT). The authors also proposed a framework for utilizing IoT sensors and machine learning algorithms to help patients create music that helps during therapy. The authors contended that such a framework could also enable remote therapy from the comfort of the patient's home.

In addition to these medical applications, sonification can be used in the IoT for safety-critical applications. For example, a smart helmet can detect harmful gases in the environment during mining operations [12] or detect gas leakage in a home [13,14]. The use of sonification has also been proposed to alert users when someone is detected in thermal imagery or other IoT sensors at critical border crossings, which can help counteract illegal border crossings [15]. Sonification in combination with IoT sensors can also prove essential in devising safety equipment for blind people. Saquib et al. [16] proposed a smart IoT device called "BlinDar", which uses ultrasonic sensors and global positioning systems (GPS) to ease navigation for blind people. GPS can also allow blind personnel to share their location with others in real time.

Sonification can also be used in combination with IoT sensors in smart city applications, such as waste collection and monitoring [17,18]. Such systems can enable efficient waste management by monitoring waste levels and can direct personnel to collect trash in high-traffic areas.

1.2. Research Problem: Modeling and Predicting Perceived and Induced Emotion

Emotions play an essential role in human behavior. Music and emotions have been studied for many years. The American Psychological Association [19] defines emotion as

"a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event".

"Affective computing" is a multidisciplinary field comprised of computer science, cognitive science, and psychology [20]. Using AI, affective computing can enable robots and computers to understand and respond to humans on a much deeper level. This intersection of AI and computer science, also called "artificial emotional intelligence", aids the development of tools for recognizing affective states and expressing emotions [21].

Affective computing enables emotion recognition in various types of multimedia, such as text, pictures, audio, and video, to create and improve user-friendly interfaces capable of parsing human emotions. Affective datasets contain lists of human annotations concerning the emotions recognized in the stimuli, which are then used to train machine learning models.

Humans can also experience emotions from music, speech, and audio files. Induced emotion refers to emotions that involve introspective perception of psychophysiological changes, whereas perceived emotion refers to listeners recognizing the emotions expressed by the external environment [22]. It is important to distinguish between induced and perceived emotions because a stimulus may invoke a different response compared to what the stimuli may actually represent. For example, listening to a cheerful song may not necessarily induce a happy emotion in the listener, despite the listener correctly perceiving the song to be a happy one.

Audio emotion recognition (AER), a subfield of emotion recognition, involves emotion recognition from music, speech, and sound events. In particular, the music industry has extensively studied the effects of soundtracks on individuals' emotions. Conventionally, emotion recognition models can be categorical or dimensional. Categorical models consider emotions with discrete labels (such as happiness, sadness, anger, fear, surprise, and disgust [23]), whereas dimensional models characterize emotions along one or more dimensions (such as arousal and valence [24]). The Geneva Emotional Music Scales (GEMS) [25] model has been widely used for measuring emotions induced by music, and the arousal–valence dimensional model has been used in studies of perceived and induced emotions [26–28].

To our knowledge, there is no comprehensive study of the performance of the prediction of perceived and induced emotions from acoustic features. In this paper, we explore emotion recognition using two datasets, IADSE [29] and Emosoundscape [30], which each represent emotions in a two-dimensional space (i.e., arousal and valence). Further, we try to identify the significant acoustic features for arousal and valence, as well as for perceived and induced emotions. The IADSE is a set of sounds for which induced emotions have been measured. The Emosoundscape dataset is a set of sounds for which perceived emotions have been measured. Analysis and modeling of these two datasets enable us to investigate and find the best models for predicting perceived and induced emotions with high accuracy.

1.3. Research Questions

This article primarily addresses the following research questions:

1. RQ1. How well do machine learning models perform when predicting arousal and valence?
2. RQ2. How different are the models that are built for predicting perceived and induced emotions?
3. RQ3. What are the significant acoustic features for predicting arousal and valence?
4. RQ4. How do the significant features vary for predicting perceived and induced emotions?

1.4. Contributions of This Work

Th purpose of this paper is to compare and contrast induced and perceived emotions from sounds with the help of various machine learning and deep learning models. We study

these two types of emotions through features that characterize different aspects of emotions. More specifically, given a set of acoustic features of sounds, the authors would like to model emotional characteristics, such as "arousal" and "valence". To build such models, the authors use two datasets, IADSE [29] and Emosoundscape [30], which are already tagged with arousal and valence. IADSE concerns induced emotions, and Emosoundscape concerns perceived emotions. We believe that the results of this research can help us further understand emotions in a better way and, thus, help in improving current IoT systems to reduce cognitive load. The key contributions of this paper are as follows:

- We present a small-scale survey of the literature related to emotion recognition, along with the features and datasets used.
- We build machine learning models to predict perceived and induced emotions.
- We compare and contrast the features used to build the best prediction models for different emotional dimensions (i.e., arousal, valence, and dominance).
- We report the significant acoustic features identified when building the best prediction models for both perceived and induced emotions.

Our results show that the machine learning models built for predicting perceived (i.e., intended) emotions are more accurate than the models built for estimating induced (i.e., felt) emotions. We also report that the accuracy of the models can be improved through acoustic feature selection, as well as by engineering and hyper-parameter tuning. Regarding the latter, machine learning techniques based on ensemble learning (e.g., Random Forests) outperform some other machine and deep learning algorithms.

This paper is organized as follows: Section 2 reviews the literature. The methodology and materials of the study are presented in Sections 3 and 4, respectively. Section 5 presents the results and analysis. Section 6 concludes the paper and highlights future research directions.

2. Related Work

The state of the art of machine learning techniques in automatic audio emotion recognition relies on characteristics of the input, output, and problem domains (types of techniques and research questions):

- Input
  - Dataset (number of samples; types of samples, e.g., sound event/music/songs/etc.)
  - Features (number of features; types of features, e.g., psycho-acoustic features, dimensions, e.g., 1D/2D)
- Output
  - Output (categorical model, e.g., sad/happy/angry/etc; dimensional model, e.g., arousal/valence/dominance)
  - Perceived or induced emotion
- Problem
  - Classification, clustering, or prediction problems
  - Evaluation metrics, e.g., RMSE/MSE/accuracy/explained variance/etc.)
  - Feature selection/reduction
  - Feature analysis (significant features, smallest number of features, and so on)

Acoustic sounds, such as music, natural, and non-speech sounds, can both elicit and convey emotions. Research concerning emotion induction has received comparatively less attention than emotion perception [22,31,32]. Perceived emotion is the emotion that the sound stimulus is intended to convey. Induced emotion is the emotion felt by the listener after introspection and processing of the sound [22,30]. Thus, perceived and induced emotions may not be the same. Table 1 shows a summary of music and audio emotion recognition in the literature.

Machine learning algorithms that perform audio emotion recognition require appropriate features to recognize emotions. Speech audio recognition using Hidden Markov

(HMM), Gaussian Mixture (GMM), and Support Vector Machine (SVM) models have categorized speech acoustic features with a high degree of accuracy [33–35]. Table 2 lists the features used for emotion recognitions in the literature.

Automatic emotion recognition in music has been a topic of interest for many researchers. The aim is to easily categorize music with similar emotions without labor-intensive human annotation. Music emotion recognition research has been conducted using regression, classification, and deep learning models.

2.1. Music Emotion Recognition

Yang et al. [36] used regression analysis to predict arousal and valence ratings found in 195 music samples that were composed of popular songs from English, Chinese, and Japanese albums. The authors reported R^2 values of 58.3% for arousal and 28.1% for valence using an SVM with 114 acoustic features, such as loudness, sharpness, and other features.

Yang and H. Chen [37] carried out an experiment to recognize emotions in music signals so that similar music could be retrieved and classified. The authors developed a custom ranking algorithm—RBF-Listnet—to optimize the retrieval of similar music samples based on the underlying emotion. The authors argued that automated retrieval reduced human annotation efforts for fetching similar music samples. The authors reported a gamma statistic of 0.326 for valence recognition.

Eerola et al. [38] proposed a model for predicting perceived emotions in a music dataset called *Soundtrack110* that contained 110 samples. The authors used a set of 29 features extracted using MIRToolbox to predict arousal and valence ratings. The authors reported an explained variance of 58% to 85% using linear regression models. The authors also reported R^2 statistics for the prediction of various categorical emotions (angry, scary, happy, sad, and tender).

Seo and Huh [39] used machine learning and deep neural networks to recognize induced emotions, with the ultimate goal being to classify similar music samples. The authors used 100 music samples from Korean pop music. The authors reported a best match rating of 73.96% via an SVM, which was slightly greater than that of the deep neural network, i.e., 72.90%.

Liu et al. [40] classified the emotions in music samples by using their spectrograms as features in a deep learning model. Spectrograms contain both time and frequency information, and the authors used them to classify similar music samples using convolutional neural networks (CNNs). The authors used a publicly available dataset called 1000-Song [41] to test the proposed model. The authors reported an average accuracy of 72.4% using the CNN model.

Fan et al. [42] proposed the use of a ranking algorithm called smoothed RankSVM (SRSVM) for ranking music with similar emotions. The authors created a corpus of 100 music clips from different musical genres. The authors utilized 56 features generated via the MIRToolbox and reported gamma statistics of of 0.801 and 0.795 for arousal and valence, respectively.

2.2. Sound Emotion Recognition

In addition to music, researchers have looked into emotion recognition with other sound stimuli, such as emotion recognition for audio samples (non-speech) that are also called *sound events* or *soundscapes*. Schafer [43] categorized soundscapes into six categories (natural sounds, human sounds, sounds and society, mechanical sounds, quiet and silence, and sounds as indicators). The categories are based on the origin of the sound source and the context in which the sound is heard [30]. Similarly to music emotion recognition, machine learning algorithms require labels to train the model to establish ground truth. Audio emotion recognition thus combines human annotation and machine learning to recognize emotions.

Schuller et al. [44] compared human annotations of emotions to those of regression with a sound dataset that contained 390 audio samples of different sounds, such as nature,

animal, and musical instrument sounds. The authors reported correlation values of 0.61 for arousal and 0.49 for valence between regression and human annotations.

Drossos et al. [45] investigated the use of rhythmic sound features for arousal prediction. The authors utilized 26 rhythm features, which were derived by applying the MIRToolbox to the IADS dataset [46]. They reported the highest accuracy of 88.37% in arousal recognition. Furthermore, feature fluctuation was found to be the best individual feature for predicting arousal values.

Fan et al. [30] created a dataset called EmoSoundscape, which contains 1213 six-second-long sounds, for soundscape emotion recognition. The authors compared the results of emotion ratings from 1182 human annotators against regression. The authors used 39 features extracted by using MIRToolbox, as well as YAAFe [47]. The authors reported the results as two protocols: A and B. Protocol A involved shuffling the sound database 10 times and then selecting sounds for training and testing (80% and 20%, respectively). Protocol B used the leave-one-out method, wherein one sound at a time was selected for training, and the remaining were used for testing during each iteration. For Protocol A, the R^2 and MSE are 0.853 and 0.049 for arousal and 0.623 and 0.128 for valence, respectively. However, for Protocol B, the R^2 and MSE were 0.855 and 0.048 for arousal and 0.629 and 0.124 for valence, respectively.

Sundaram and R. Schleicher [48] developed an audio-based retrieval system to retrieve similar sounds by querying the system. The authors selected sounds from the BBC sound effects library (http://bbcsfx.acropolis.org.uk (accessed on 1 July 2021)) and the IADS dataset to build the system. The authors also collected human annotations for these sounds to compare them against the emotional ratings of the sounds retrieved by the system. For each query, the system retrieved the top five similar sounds by using MFCC features with similar features in the latent space. The average RMSE between the queried and retrieved sounds was found to be between 1.2 to 2.6.

Researchers have also used neural networks and deep neural networks for predicting emotions in sounds. Fan et al. [49] evaluated the use of deep learning models, such as CNNs and Long Short-Term Recurrent Neural Networks (LSTM-RNNs), for sound emotion recognition using the EmoSoundscapes [30] dataset. The authors compared the performance of five deep learning architectures in predicting arousal and valence ratings. The authors used two sets of techniques to extract features. The first method used a pre-trained deep neural network created by S. Hershey et al. [50], whereas the second method involved 54 features extracted using MIRToolbox and YAAFE. The best performance for arousal was reported with the CNN with an R^2 and MSE of 0.832 and 0.035, respectively, whereas the best performance for valence was reported to have an R^2 of 0.759 and MSE of 0.078 via VGGish (a deep CNN model). The authors also investigated the arousal and valence prediction for various sounds using Schafer's categories.

Ntalampiras and Potamitis [51] used a deep learning model called the echo state network to study the similarities between music and sound datasets in eliciting emotions. The authors used three feature sets—Mel-Spectrum (MFCC), temporal modulation, and Perceptual Wavelet Packets (PWP)—and each was extracted from the IADS and 10,000 song datasets. The authors first trained the network on the music dataset and then used that trained network for sounds in the IADS dataset to determine if the arousal–valence prediction would improve. The best performance was achieved using the temporal modulation features with a mean square error of 3.13 for arousal and 3.10 for valence when using GMM clustering as the regressor.

Ntalampiras [52] compared emotion prediction using two CNNs that were designed to individually predict arousal and valence. The authors used the EmoSoundscapes data. They extracted features by employing a sample window of the audio files and then applying Fourier transformation to yield 23 features that were similar to the MFCC obtained from MIRToolbox. The authors reported an MSE value of 0.0168 for arousal and 0.0107 for valence. The authors also predicted arousal–valence ratings for sound categories as per Schafer's taxonomy.

Cunningham et al. [53,54] used shallow neural networks and regression to predict emotion using the IADS dataset. The authors employed 76 MFCC features using the MIRToolbox. The authors reported an RMSE of 0.989 and an R^2 of 0.28 for arousal, as well as an RMSE of 1.645 and R^2 of 0.12 for valence by using regression. However, an arousal with an RMSE of 0.987 and R^2 0.345 and valence with an RMSE of 0.514 and R^2 of 0.269 were achieved by using a neural network.

Researchers have also studied the effect of manipulation of sound on arousal–valence emotion prediction. Drossos et al. [55,56] created a sound dataset called BEADS, which contained binaural sound clips. The dataset is publicly available and consists of 32 sounds annotated with emotion labels. These sounds have been adjusted across five spatial positions (0, 45, 90, 135, and 180 degrees). The authors also reported a comparison of BEADS with the IADS dataset. The authors observed maximum arousal differences of 2.47 and 2.07 for arousal and valence between IADS and BEADS, respectively. Additionally, sounds at a 0 degree spatial angle elicited a higher arousal rating and a lower valence rating than those at other angles. Asutay et al. [57] conducted an experimental study to understand whether distorting the sound to reduce its identifiability caused any changes in the perceived emotions. Three different studies with participants were undertaken. Participants in each study rated both the distorted and original sound recordings from the IADS dataset using the Self-Assessment Manikin (SAM) scale; the recordings were either introduced one after the other or in a random order [58]. The third group (i.e., the control group) was presented with the original sounds and their textual descriptions before being asked to rate them. The authors contended that the processed sounds were emotionally neutral, but the participants were still able to identify them with the help of priming. Thus, the authors argued that sound designers should focus not just on the physical properties of the sounds, but also on psycho-acoustical features in order to evoke the desired emotions.

Table 1. Summary of emotion recognition in the literature.

Ref.	Problem Formulation	Emotion		Results
		Cat./Dim.	I/P	
[36]	Regression analysis	Dim. (Ar.–Val.)	P	R^2 stats.: 58.3% Ar., 28.1% Val.
[37]	Ranking	Dim. (Ar.–Val.)	P	Gamma stats.: 0.326 val.
[38]	–	Dim. and Cat. (5 emot.)	P	R^2 stats.: 77% Ar., 70% Val.
[39]	–	Dim. and Cat. (4 quadrants)	I	Acc.: 73.96% (SVM)
[40]	Classification	Dim. (Ar.–Val.)	P	Acc.: 72.4% (CNN)
[42]	Ranking	Dim. (Ar.–Val.)	P	Gamma stats.: 0.801 Ar., 0.795 Val.
[44]	Regression	Dim. (Ar.–Val.)	P	Corr. Coeff. 0.61 Ar., 0.49 Val.
[45]	Classification and Ranking	Dim. (Ar.)	P	Acc.: 81.44% Ar. (Log. regr.)
[55,56]	Annotator Ratings	Dim. (Ar.–Val.)	P	Diff. in Mean Ar.: 0.18, Mean Val.: 0.38 IADS and BEADS
[30]	Regression	Dim. (Ar.-Val.)	P	R^2: 0.853 Ar., 0.623 Val.
[49]	Classification and Regression	Dim. (Ar.–Val.) and Cat. (Schafer's)	P	R^2: 0.892 Ar., 0.759 Val.
[51]	Regression, Clustering	Dim. (Ar.–Val.)	P	MSE: 3.13 Ar., 3.10 Val.
[52]	Prediction	Dim. (Ar.–Val.) and Cat. (Schafer's)	P	MSE: 0.0107 Ar., 0.0168 Val.
[53,54]	Regression	Dim. (Ar.–Val.)	P	R^2: 0.345 Ar., 0.269 Val.
[57]	Annotator Ratings	Dim. (Ar.,Val., Annoyance, and Loudness)	I	–
[48]	Latent Analysis and Retrieval	Dim. (Ar.–Val.–Domn.)	P	RMSE for top-5 clips b.w. 1.6 to 2.6

Abbreviations: Ar.—Arousal, Val.—Valence, I—Induced, P—Perceived, Cat.—Categorical, Dim.—Dimensional.

Table 2. Summary of features used for emotion recognition in the literature.

Ref.	Dataset	Features	
		No. of Features	Type
[36]	195 Pop Songs	114	PsySound, Spectral Contrast, Daubechies wavelets coefficient histogram (DWCH)
[37]	60 K-pop and 1240 Chinese pop music samples	157	10 Melody, 142 timbre, 5 rhythm
[38]	SoundTrack110-110 song samples	29	dynamics, timbre, harmony, register, rhythm, articulation
[39]	100 K-pop songs		Avg. height, peak avg., HfW, avg. width, BPM
[40]	1000 song dataset [41]	30,498	Spectrograms
[42]	100 Emusicclips	56	features of MIR Toolbox [59]
[44]	Emotional Sound Database (390 sounds)	73	31 low-level descriptors (energy, spectral, and voicing) and 42 functional (statistical, regression, and local minima/maxima)
[45]	IADS dataset (167 sounds) [46]	26	beat spectrum, onsets, tempo, fluctuation, even density, and pulse clarity
[55,56]	Binaural sound corpus—BEADS (167 sounds)	5	Angular adjustments (45°, 90°, 135°, and 180°)
[30]	EmoSoundscapes—1213 soundscape files	39	features of MIRToolbox and YAAFE
[49]	*EmoSoundscapes* [30]	54	loudness, MFCC, energy, spectral
[51]	IADS [46] and 1000 songs [41]		MFCC and Perceptual Wavelet Packets (PWP)
[52]	*EmoSoundscapes* [30]	23	MFCC-like
[53,54]	IADS [46]	76	Features of MIRToolbox
[57]	18 envir. sounds from IADS [46]		Fourier-time transformation (FTT)
[48]	2491 audio clips from the BBC Sound Effects Library	12	MFCC

3. Experimental Setup

3.1. Datasets and Psychoacoustic Features

To conduct our experiment, we utilized two datasets: IADSE [29] and EmoSoundscape [30]. These datasets contain sound samples with their annotated emotions. EmoSoundscape contains ratings for perceived emotions and uses a two-dimensional space (arousal/valence). IADSE contains ratings for induced emotions and uses a three-dimensional space (arousal/valence/dominance). Therefore, we chose these two datasets to compare induced and perceived emotion predictions.

To extract the features from these datasets, we used the MIRToolbox [59], which extracts (psycho)acoustic and musically related features from databases of audio files for statistical analysis [59]. Following Lange and Frieler [60], a total of 68 features were extracted from each stimulus to represent either the arithmetic mean or the sample standard deviation of the frame-based features computed over default window sizes (typically 50 ms for low-level features and 2–3 s for medium-level features) and a 50% overlap. The selected features represent the following families:

- *Dynamics*—intensity of the signal, such as the root mean square (RMS) of the amplitude;
- *Rhythm*—articulation, density, and temporal periodicity of events, such as the number of events per second (event density);
- *Timbre/Spectrum*—brightness, noisiness, dissonance, and shape of the frequency spectrum, such as the spectral center of mass (centroid);

– *Pitch*—presence of harmonic sounds, such as the proportion of frequencies that are not multiples of the fundamental frequency (inharmonicity);
– *Tonality*—presence of harmonic sounds that collectively imply a major or minor key, such as the strength of a tonal center (key clarity).

Although most features represent relatively low-level acoustic or auditory attributes (e.g., RMS), some are based on perceptual models (e.g., roughness), and yet others are based on cognitive models that presume long-term exposure to the stimulus domain (e.g., key clarity). A summary of the features is shown in Table 3.

Table 3. Number of acoustic features captured by the MIRToolbox.

Selected MIR Features	
Feature	**Count**
Dynamics	2
Pitch	1
Rhythm	6
Spectral	23
Spectral MFCC	26
Timbre	4
Tonal	6
Total	68

We computed the pairwise correlation for the 68 features in each dataset. The seven pairs of features shown in Table 4 are the features in the EmoSoundscape dataset with correlations greater than 90%. In addition, the four pairs of features shown in Table 5 are the features in the IADSE dataset with correlations greater than 90%.

Table 4. Highly correlated features in the EmoSoundscape dataset.

Feature 1	Feature 2	Correlation
timbre spectralflux (std)	dynamics rms (std)	0.954
spectral spread (mean)	spectral rolloff95 (mean)	0.949
timbre spectralflux (mean)	dynamics rms (mean)	0.946
spectral rolloff85 (mean)	spectral centroid (mean)	0.94
spectral skewness (mean)	spectral kurtosis (mean)	0.93
spectral mfcc 12 (std)	spectral mfcc 11 (std)	0.905
spectral rolloff85 (mean)	spectral flatness (mean)	0.902

Table 5. Highly correlated features in the IADSE dataset.

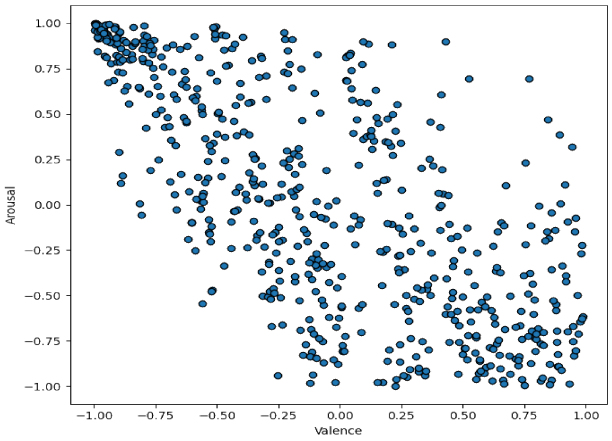
Feature 1	Feature 2	Correlation
spectral kurtosis (mean)	spectral skewness (mean)	0.968
spectral rolloff85 (mean)	spectral centroid (mean)	0.951
spectral rolloff95 (mean)	spectral spread (mean)	0.927
spectral rolloff85 (mean)	spectral rolloff95 (mean)	0.923

3.1.1. EmoSoundscape Dataset: A Dataset for "Perceived" Emotion

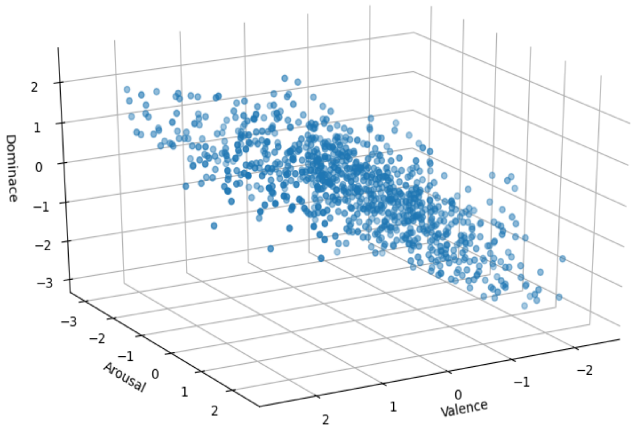
We used the EmoSoundscape dataset [30], which consists of two subsets. The first subset contains 600 audio samples categorized into 6 groups of 100 samples each according to

Schafer's soundscape taxonomy; these groups are natural sounds, human sounds, sounds and society, mechanical sounds, quiet and silence, and sounds as indicators. The second subset contains 613 samples; each is a mix of soundscapes from two or three of the first subset's classes. All of these soundscapes are annotated with their perceived emotion, including arousal and valence. The first subset of this dataset was used for our experiment.

Because the EmoSoundscape dataset contains arousal and valence for each sound sample, the scatter plot of valence versus arousal for the EmoSoundscape dataset is shown in Figure 1a. Note that both variables are z-normalized.



(**a**)



(**b**)

Figure 1. Visualization of EmoSoundscape and IADSE data sets. (**a**) Scatter plot of the normalized data points of EmoSoundscape in the AV space. (**b**) Scatter plot of normalized data points of IADSE in the AVD space.

3.1.2. IADSE Dataset: A Dataset for "Induced" Emotion

We also used the IADSE dataset, which contains 935 sounds, with each sound rated by at least 100 listeners on 9-point Likert scales for the dimensions of felt arousal, valence, and dominance (induced emotions). In addition, a scatter plot of the data points in the IADSE dataset is shown in Figure 1b.

3.2. Evaluation Metrics for Analysis

To measure the performance of the regression models, different common metrics can be utilized, including the mean absolute error (MAE), mean squared error (MSE), root mean square error ($RMSE$), R-squared (R^2), median absolute error, max error, and explained variance. These evaluation metrics have been heavily utilized in the machine learning

literature and they are the main evaluation metrics in the context of evaluating machine learning models. $RMSE$, MSE, and R^2 were chosen to evaluate the performance of each regression model. It is important to note that the prime objective of this study was to compare the performance of various machine learning models in predicting emotion dimensions (i.e., valence, arousal, dominance) and not to conduct controlled experiments and perform statistical significance tests.

The mean square error (MSE) is the average of the square of the errors; the larger the value is, the larger the error will be.

$$MSE = \frac{\Sigma(y_i - \hat{y}_i)^2}{n}$$

The root mean square error ($RMSE$) can be considered as the standard deviation of the prediction errors. Because it applies a high penalty for large errors, it is beneficial when large errors are undesirable.

$$RMSE = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n}}$$

Finally, R^2 represents the ratio of the total sum of squares of the prediction error to the total sum of squares of error with the mean; the closer the value of R^2 is to 1, the better the regression model will be.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y}_i)^2}$$

It should be mentioned that R^2 is a less commonly used metric for assessing non-linear models [61].

4. Methodology

In our previous work with the IADS and the EmoSoundscape datasets [62], we reported that Random Forest outperformed other models in A/V prediction using a 1D psycho(acoustic) feature set, while other models mostly suffered from overfitting. This result is somewhat expected because ensemble models reduce the risk of overfitting. Ensemble models combine the prediction results of several base models. In addition, among ensemble models, Random Forest is preferable for overfitting problems [63]. Unlike other ensemble methods, adding more trees in a Random Forest model does not increase the risk of overfitting. Therefore, we chose Random Forest as one of the prediction models for these datasets in this article. Random Forest (RF) is an ensemble method that averages the prediction results of several decision trees. To compare the prediction results from the ensemble model (RF) with deep models, we developed a multilayer perceptron model and a 1D convolutional neural network model. For all of the models, we used 30% of the data as the test data and also applied 5-fold cross validation (CV). In order to compute the training and testing errors, we averaged the RMSE values over these 5 folds.

4.1. Feature Selection

Feature selection techniques can be divided into two main types. The *filter* methods are usually performed as a preprocessing step by using the underlying properties of the features measured with different univariate statistics. The other method uses an estimator to perform feature selection, so it is considered as a *wrapper* method. It selects the features based on the performance of a model. The filter-based methods are faster, whereas, the wrapper methods are more computationally expensive.

In our previous work [62], we used a filter-based method called the "univariate linear regression test" (KBest) for selecting the k best features. In contrast, in this work, we used a wrapper method called Recursive Feature Elimination (RFE) [64]. Using RFE, we applied Random Forest as the estimator to be fitted to the datasets. Then, the features were ranked based on their weights, and the features with the lowest weight were removed. This process was repeated until a desirable number of features remained. Because we did not have prior

knowledge about the number of best features, we tried the number of features as a value ranging from 1 to 68 in the hyper-parameter tuning phase, along with selecting the best parameters for each Random Forest regressor.

4.2. Hyper-Parameter Tuning

Hyper-parameter tuning is the process of selecting the best parameters for a model to obtain the optimal results. Grid search is a technique that can be employed to find the optimal parameters of the model through which all combinations of the determined values for parameters are examined. We performed an exhaustive search for hyper-parameter tuning on 90 parameters overall, including 22 Random Forest parameters and 68 RFE parameters, in order to find the optimal values. Here is the list of parameters that were tuned:

- $n\_estimators : (50, 100, 150, 200, 250, 300)$, number of trees in the forests;
- $max\_depth : (5, 10, 20, 30, 50)$, maximum number of levels in each decision tree;
- $min\_samples\_split : (2, 3, 4, 5, 6, 7)$, minimum number of data points placed in a node before the node is split;
- $min\_samples\_leaf : (1, 2, 3, 5)$, minimum number of data points allowed in a leaf node;
- $k : range(1, 68)$, number of features selected using RFE with the RF estimator.

5. Results and Analysis

This section reports the performance and the results of the analysis by comparing predictions for both perceived and induced emotions.

5.1. Performance of Prediction Models

The arousal, valence, and dominance predictions on the IADSE dataset and the arousal and valence predictions on the EmoSoundscape dataset using the tuned Random Forest models are shown in Table 6.

Table 6. The results of our previous and current work.

Our Previous Work		**Arousal**			**Valence**		
		EmoSoundscape	**IADS**	**IADSE**	**EmoSoundscape**	**IADS**	**IADSE**
DS	No. of Samples	600	167	927	600	167	927
KBest	No. of Features	26/68	28/313	27/68	29/68	27/313	27/68
Random Forest Hyper-parameters	No. of Estimators	200	200	30	100	150	150
	Max Depth	20	10	20	30	5	20
	Min Samples Split	3	5	4	2	2	2
	Min Samples Leaf	1	2	1	1	2	1
Evaluation Metrics	Train RMSE	0.09	0.36	0.30	0.013	0.41	0.41
	Test RMSE	0.25	0.88	0.78	0.37	0.98	1.13

Our Current Work		**Arousal**		**Valence**		**Dominance**
		EmoSoundscape	**IADSE**	**EmoSoundscape**	**IADSE**	**IADSE**
DS	No. of Samples	600	927	600	927	927
RFE	No. of Features	15/68	25/68	14/68	9/68	7/68
Random Forest Hyper-parameters	No. of Estimators	50	300	50	250	150
	Max Depth	20	20	10	30	5
	Min Samples Split	5	3	5	2	2
	Min Samples Leaf	2	1	2	1	5

Table 6. *Cont.*

Evaluation Metrics	Train RMSE	0.1032	0.2818	0.1681	0.3970	0.70
	Test RMSE	0.2351	0.7782	0.3698	1.1577	0.83
	Train MSE	0.0106	0.0794	0.0283	0.1576	0.49
	Test MSE	0.0552	0.6055	0.1367	1.3402	0.70
	Train R2	0.9718	0.9422	0.9137	0.9200	0.48
	Test R2	0.8639	0.5631	0.5860	0.3700	0.26

The best arousal prediction for the EmoSoundscape dataset was achieved with 15 features. The best evaluation metrics were 0.24, 0.05, and 0.86 for the test $RMSE$, MSE, and R^2, respectively. On the other hand, for the IADSE dataset, the best arousal prediction was achieved with 25 features. The best evaluation metrics were 0.78, 0.61, and 0.56 for the $RMSE$, MSE, and R^2, respectively.

Regarding valence, the best prediction on the EmoSoundscape dataset was achieved using 14 features. The best evaluation metrics were 0.37, 0.14, and 0.59 for the test $RMSE$, MSE, and R^2, respectively. On the other hand, for the IADSE dataset, the best valence prediction was achieved with nine features. The best evaluation metrics were 1.16, 1.34, and 0.37 for the test $RMSE$, MSE, and R^2, respectively. For the dominance prediction, we achieved 0.83, 0.70, and 0.26 for the test $RMSE$, MSE, and R^2, respectively, using seven features.

To compare the performance of the RF model as an ensemble method with deep models, two deep neural networks were developed to predict perceived and induced emotions. Specifically, a four-layer perceptron and a one-layer convolutional neural network followed by a four-layer perceptron were utilized. These models fit the data using all features and the selected features that were identified by the exhaustive search. Table 7 shows the emotion predictions using these deep models.

Table 7. Comparing the ensemble model with deep models.

		RMSE—IADSE						**RMSE—EmoSoundscape**			
		Arousal		**Valence**		**Dominance**		**Arousal**		**Valence**	
Model	**Features**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Tuned RF	all	–	–	–	–	–	–	–	–	–	–
	selected	0.28	**0.77**	0.39	**1.15**	0.70	**0.83**	0.10	**0.23**	0.16	**0.36**
4L MLP	all	0.22	0.83	0.22	1.23	0.20	0.89	0.06	0.26	0.06	0.39
	selected	0.30	0.81	0.60	1.19	0.55	0.89	0.09	0.25	0.13	0.39
1D CNN	all	0.85	0.89	1.41	1.44	0.97	0.97	0.29	0.31	0.41	0.42
	selected	0.83	0.87	1.27	1.37	0.97	0.97	0.24	0.26	0.38	0.39
Average	all	0.645	0.860	0.925	1.335	0.585	0.93	0.175	0.285	0.235	0.405
	selected	0.476	0.816	0.753	1.236	0.74	0.916	0.143	0.246	0.223	0.380

Although the performance of the deep models was close to performance of the tuned RF, RF achieved a better emotion prediction for both datasets.The best test errors using tuned RF are shown in bold in Table 7. In addition, in most cases, the performance of the deep models using selected features was better than their performance using all features, which indicates the effectiveness of our selected features in predicting emotions in each dataset.

To compare the performance of our work with that of similar works in the literature, a few papers reporting their arousal and valence prediction results on the EmoSoudscape dataset were found. Fan et al. [42] reported MSE values of 0.049 and 0.128 for predicting arousal and valence, respectively. Converting their results into RMSE, they achieved 0.22 and 0.36 for arousal and valence prediction using a Support Vector Regressor (SVR), which

were close to our results obtained using the tuned Random Forest. Furthermore, they improved their results by augmenting the dataset and applying a tuned convolutional neural network (CNN). Since their work was on the augmented dataset, their results are not comparable with ours.

Part of the work performed by Ntalampiras [52] was on the EmoSoundscape dataset, and they used CNN models. The MSE values reported for arousal and valence prediction were around 0.049 and 0.11, respectively, which are equivalent to 0.22 and 0.33 for RMSE. These results are close to other reported performances on the EmoSoundscape dataset, with a slight improvement in valence prediction.

Ntalampiras [52] also applied a CNN on the data collected from both subsets of the EmoSoundscape dataset and achieved better performance for both arousal and valence. It must be noted that the second subset of EmoSoundscape contains sound events that are mix of soundscapes from two or three of the sound events in the first subset. We could not identify any work on the IADSE dataset with which to compare our results.

With respect to the research question RQ1 and according to Table 7, we observe that the selected machine learning prediction models have similar performance, except that the model built based on the optimized Random Forests outperforms the other models in reducing the RMSE values for both arousal and valence. Regarding RQ2, Table 7 indicates that the RMSE models built for predicting perceived emotions (i.e., EmoSoundscape) are associated with lower RMSE values in comparison with the RMSE values captured for models predicting induced emotions (i.e., IADSE). The average RMSE values for both the training and testing datasets computed for induced emotion dimensions (i.e., arousal and valence) are substantially low compared to the RMSE values captured for perceived emotion dimensions. This observation implies that modeling induced emotion is harder than building models for predicting perceived emotion.

5.2. Significant Features

The significant features used in the tuned Random Forest models for each prediction of emotion for both datasets are sorted in Tables 8 and 9. In addition, Table 10 provides a better insight about the common features among different emotion predictions and among/within the IADSE and EmoSoundscape datasets. In Table 10, significant features for emotion prediction using the tuned Random Forest are indicated by '*' in the IADSE dataset and by '+' in the EmoSoundscape dataset. Furthermore, if any of these significant features have highly correlated features in their peer datasets, these correlated features are marked as (*) and (+) for the IADSE and EmoSoundscape datasets, respectively.

Table 8. Sorted significant features for arousal (15 features) and valence (14 features) by using RFE and applying the RF model on the EmoSoundscape dataset.

Arousal 15/68	Sign.
spectral roughness (mean)	0.704
timbre spectralflux (mean)	0.085
dynamics rms (mean)	0.029
spectral brightness (mean)	0.027
spectral spectentropy (mean)	0.023
spectral rolloff85 (mean)	0.018
spectral rolloff95 (mean)	0.015
rhythm fluctuationmax peakposmean	0.015
spectral mfcc 13 (std)	0.014
spectral mfcc 12 (std)	0.012
spectral irregularity (mean)	0.010

Table 8. *Cont.*

spectral mfcc 5 (mean)	0.010
timbre lowenergy (std)	0.010
tonal hcdf (mean)	0.010
rhythm attacktime (mean)	0.009
Valence 14/68	**Sign.**
spectral roughness (mean)	0.374
dynamics rms (mean)	0.155
timbre lowenergy (std)	0.105
spectral mfcc 6 (std)	0.054
spectral centroid (std)	0.043
spectral mfcc 4 (std)	0.042
rhythm pulseclarity (mean)	0.033
spectral skewness (mean)	0.032
spectral mfcc 9 (std)	0.031
spectral mfcc 8 (mean)	0.029
spectral rolloff95 (mean)	0.026
spectral flatness (std)	0.024
spectral rolloff95 (std)	0.024
timbre spectralflux (mean)	0.020

Table 9. Sorted significant features for arousal (25 features), valence (nine features), and dominance (seven features) by using RFE and applying the RF model on the IADSE dataset.

Arousal	**Sign.**
timbre spectralflux (mean)	0.289
dynamics rms (mean)	0.091
dynamics rms (std)	0.064
spectral flatness (std)	0.044
spectral roughness (mean)	0.041
spectral spectentropy (mean)	0.036
spectral rolloff85 (mean)	0.035
spectral brightness (mean)	0.032
rhythm pulseclarity (mean)	0.031
timbre spectralflux (std)	0.029
tonal keyclarity (std)	0.024
spectral skewness (mean)	0.024
rhythm tempo (std)	0.024
pitch pitch (mean)	0.022
spectral flatness (mean)	0.021
timbre lowenergy (mean)	0.020
tonal keyclarity (mean)	0.019
spectral mfcc 6 (mean)	0.018
spectral centroid (mean)	0.018
spectral spectentropy (std)	0.018

Table 9. *Cont.*

spectral brightness (std)	0.018
spectral spread (mean)	0.018
spectral irregularity (mean)	0.018
spectral mfcc 8 (mean)	0.017
spectral mfcc 2 (mean)	0.016
Valence	**Sign.**
tonal keyclarity (mean)	0.248
spectral roughness (mean)	0.136
spectral rolloff85 (std)	0.101
dynamics rms (std)	0.092
rhythm fluctuationmax peakposmean	0.092
spectral brightness (mean)	0.092
spectral spread (mean)	0.083
spectral skewness (std)	0.082
spectral mfcc 11 (std)	0.070
Dominance	**Sign.**
spectral roughness (mean)	0.367
dynamics rms (mean)	0.158
tonal keyclarity (mean)	0.115
tonal hcdf (std)	0.111
timbre spectralflux (mean)	0.091
spectral brightness (mean)	0.089
spectral mfcc 3 (mean)	0.065

Table 10. Significant features for each prediction in the IADSE and EmoSoundscape datasets that showed common features among them. If any of these significant features had highly correlated features in its peer dataset, these correlated features were marked as (*) and (+) for the IADSE and EmoSoundscape datasets, respectively.

Features	**IADSE**			**EmoSoundscape**	
Dynamics	Arousal	Valence	Dominance	Arousal	Valence
dynamics rms (mean)	*		*	+	+
dynamics rms (std)	*	*			
Pitch					
pitch pitch (mean)	*				
Rythm					
rhythm pulseclarity (mean)	*				+
rhythm tempo (std)	*				
rhythm fluctuationmax peakposmean		*		+	
rhythm attacktime (mean)				+	
Timber					
timbre spectralflux (mean)	*		*	+	+
timbre spectralflux (std)	*				
timbre lowenergy (mean)	*				
timbre lowenergy (std)				+	+

Table 10. *Cont.*

Features	IADSE			EmoSoundscape	
Tonal					
tonal keyclarity (mean)	*	*	*		
tonal keyclarity (std)	*				
tonal hcdf (mean)				+	
tonal hcdf (std)			*		
Spectral					
spectral flatness (std)	*				
spectral roughness (mean)	*	*	*	+	+
spectral spectentropy (mean)	*			+	
spectral rolloff85 (mean)	*			+	
spectral rolloff85 (std)		*			
spectral brightness (mean)	*	*	*	+	
spectral brightness (std)	*				
spectral skewness (mean)	*				+
spectral skewness (std)		*			
spectral flatness (mean)	*			(+)	
spectral flatness (std)					+
spectral centroid (mean)	*			(+)	
spectral centroid (std)					+
spectral spectentropy (std)	*				
spectral spread (mean)	*	*		(+)	(+)
spectral irregularity (mean)	*			+	
spectral rolloff95 (mean)	(*)	(*)		+	+
spectral rolloff95 (std)					+
spectral kurtosis (mean)					(+)
Spectral-mfcc					
spectral mfcc 5 (mean)				+	
spectral mfcc 6 (mean)	*				
spectral mfcc 8 (mean)	*				+
spectral mfcc 2 (mean)	*				
spectral mfcc 3 (mean)			*		
spectral mfcc 4 (std)					+
spectral mfcc 6 (std)					+
spectral mfcc 9 (std)					+
spectral mfcc 11 (std)		*		(+)	
spectral mfcc 12 (std)				+	
spectral mfcc 13 (std)				+	

Considering the induced emotion predictions for the IADSE dataset, 25 features were considered as the significant features for predicting induced arousal, whereas nine features were considered for induced valence. There were five common features for predicting induced arousal and valence: (1) dynamics rms (std), (2) tonal keyclarity (mean), (3) spectral roughness (mean), (4) spectral brightness (mean), and (5) spectral spread (mean). We can also consider spectral rolloff95 (mean) as the sixth common feature because it had a high correlation with spectral spread (mean) in the IADSE dataset. Furthermore, for the

dominance prediction, seven significant features were identified, and three of these features were indicated as significant features for the induced arousal and valence predictions.

On the other hand, in predicting perceived emotions in the EmoSoundscape dataset, 15 features were considered as the significant features for predicting perceived arousal, whereas 14 features were considered for perceived valence. There were five common features for predicting induced arousal and valence: (1) dynamics rms (mean), (2) timbre lowenergy (std), (3) timbre spectralflux (mean), (4) spectral roughness (mean), and (5) spectral rolloff95 (mean). We can also treat spectral spread (mean) as the sixth common feature because it has a high correlation with spectral rolloff95 (mean) in the EmoSoundscape dataset.

Comparing the arousal and valence predictions based on these two datasets, 25 features were considered as the significant features for predicting induced arousal (IADSE), whereas 15 features were considered for perceived arousal (EmoSoundscape). There were seven common features for predicting induced and perceived arousal: (1) dynamics rms (mean), (2) timbre spectralflux (mean), (3) spectral roughness (mean), (4) spectral rolloff85 (mean), (5) spectral irregularity (mean), (6) spectral spectentropy (mean), and (7) spectral brightness (mean). We can also treat spectral centroid (mean) as the eighth common feature because it had a high correlation with spectral rolloff85 (mean) in the EmoSoundscape dataset.

In addition, 15 features were considered as the significant features for predicting induced valence (IADSE), whereas nine features were considered for perceived valence (EmoSoundscape). There was only one common feature for predicting induced and perceived arousal, which was spectral roughness (mean). We can also treat spectral spread (mean) and spectral rolloff95 (mean) as the second and third common features because these two features demonstrated a high correlation with each other in both datasets, and each of them was selected as a significant feature for predicting induced and perceived valence.

With respect to RQ3 and according to Tables 8 and 9, we observe that the number of significant features for predicting arousal is greater than the number of significant features for predicting valence. This observation implies that predicting arousal-based emotions, such as excitement, is harder than predicting valence-based emotions, such as positiveness. Therefore, to model these two dimensions, different numbers and sets of features would be required. With respect to RQ4 and according to Table 8 for perceived emotions and Table 9 for induced emotions, we observe that the number of significant features for predicting arousal in induced emotion (i.e., the IADSE dataset in Table 9) is substantially greater than the number of significant features listed for arousal prediction in perceived emotion (i.e., EmoSoundscape dataset in Table 9). In an analogous way, this observation indicates that modeling induced emotions is substantially harder than building models for predicting perceived emotions.

6. Conclusions and Future Work

It is important to monitor devices and their activities when connected to a network and to detect possible threats or events that occur in the system. In addition to conventional communication channels, such as textual descriptions and visualization, sonification is an effective technique for quickly directing users' attention to interconnected devices [65,66]. One of the major advantages of using sounds rather than textual descriptions and visualizations to alert users is that operators can listen to sounds and use visual displays at the same time without significantly increasing cognitive workload.

Although informative communication through sounds (i.e., sonification) is very promising, implementing sonification comes with its own challenges and problems. One of the key challenges is the design of proper and representative sounds for specific events. There are several issues when designing a sound to represent an event. The issues include if the sound should represent the semantics and meaning of the event, it needs to point out spatial information that is needed to trace the events, or it needs to convey the impact of each event to the user. Furthermore, the sounds selected for sonifying these events or

semantics need to be tested for their usability in order to find out whether they convey the required information.

In addition to the above issues, the psychological impact of each sound is also a key issue when designing sonifications for a large and complex system, such as the IoT. More specifically, it is important to have a clear understanding of the impact of each sound on users. As such, it is important to understand perceived (i.e., expressed emotion) and induced (i.e., felt) emotions.

This paper investigates whether it is possible to build machine-learning-based models to predict perceived and induced emotion, where emotion is defined based on three dimensions: (1) arousal, (2) valence, and (3) dominance. To perform the research and analysis, we utilized two datasets—one that concerns perceived emotions and another that concerns induced emotions. The EmoSoundscape dataset measures a user's perceived emotion, whereas the IADSE dataset quantifies a user's induced emotion. Our initial assumption was that it would be more difficult to model and predict induced emotion in comparison with perceived emotion.

Our findings confirm our assumption in that it is relatively more difficult to predict induced emotion than perceived emotion. As highlighted in Table 7, the RMSE values obtained for training and testing of models built for the IADSE (i.e., induced emotion) are greater than those calculated for the EmoSoundscape dataset (i.e., perceived emotion). We also observed that the models built for both induced and perceived emotion are of moderate accuracy, which indicates that identifying the optimal and best models for predicting these emotions is generally a difficult task.

The research reported in this paper needs further improvement and more comprehensive analysis. In particular, due to the great performance of ensemble learning approaches (more specifically, Random Forests), some other ensemble-learning-based approaches need to be explored with the intention of optimizing the best models. We also need to conduct additional research with perceived and induced emotion in certain contexts, such as security and monitoring of the IoT, and to determine if certain emotions cause the operator of a system to react in a certain and robust way. More precisely, it is important to understand how perceived and induced emotions trigger actions that we expect when certain events occur in the IoT.

Author Contributions: Data curation, D.R.W.S.; Formal analysis, F.A.; Funding acquisition, A.S.N. and K.S.J.; Methodology, F.A. and L.F.G.; Supervision, A.S.N.; Writing—original draft, F.A. and P.D.; Writing—review and editing, D.R.W.S., A.S.N. and K.S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Science Foundation (NSF) under grant numbers CNS-1347521 and SES-1564293.

Data Availability Statement: Data will be available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Real-Time IoT Monitoring—Visualize Device Performance. Available online: https://www.datadoghq.com/ (accessed on 1 July 2021).
2. Khan, W.; Ansell, D.; Kuru, K.; Bilal, M. Flight guardian: Autonomous flight safety improvement by monitoring aircraft cockpit instruments. *J. Aerosp. Inf. Syst.* **2018**, *15*, 203–214.
3. Saraubon, K.; Anurugsa, K.; Kongsakpaibul, A. A Smart System for Elderly Care Using IoT and Mobile Technologies. In Proceedings of the ICSEB '18—2018 2nd International Conference on Software and E-Business, Zhuhai, China, 18–20 December 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 59–63. [CrossRef]
4. Sainadh, A.V.M.S.; Mohanty, J.S.; Teja, G.V.; Bhogal, R.K. IoT Enabled Real-Time Remote Health Monitoring System. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; pp. 428–433. [CrossRef]
5. Shahada, S.A.A.; Hreiji, S.M.; Atudu, S.I.; Shamsudheen, S. Multilayer Neural Network Based Fall Alert System Using IOT. *Int. J. MC Sq. Sci. Res.* **2019**, *11*, 1–15. [CrossRef]

6. Mwangi, A.; Ndashimye, E.; Karikumutima, B.; Ray, S.K. An IoT-alert System for Chronic Asthma Patients. In Proceedings of the 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 4–7 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 12–19.
7. Danna, J.; Velay, J.L. Handwriting Movement Sonification: Why and How? *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 299–303.
8. Turchet, L. Interactive sonification and the IoT: The case of smart sonic shoes for clinical applications. In Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound, Nottingham, UK, 18–20 September 2019; pp. 252–255. [CrossRef]
9. Rutkowski, T.M. Multichannel EEG sonification with ambisonics spatial sound environment. In Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, Siem Reap, Cambodia, 9–12 December 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1–4.
10. Quasim, M.T.; Alkhammash, E.H.; Khan, M.A.; Hadjouni, M. Emotion-based music recommendation and classification using machine learning with IoT Framework. *Soft Comput.* **2021**, *25*, 12249–12260.
11. Timoney, J.; Yaseen, A.; Mcevoy, D. The Potential Role of Internet of Musical Things in Therapeutic Applications. In Proceedings of the 10th Workshop on Ubiquitous Music (UbiMus 2020), g-ubimus, Porto Seguro, BA, Brazil, 5–7 August 2020. [CrossRef]
12. Roja, P.; Srihari, D. Iot based smart helmet for air quality used for the mining industry. *Int. J. Res. Sci. Eng. Technol.* **2018**, *4*, 514–521.
13. Meshram, P.; Shukla, N.; Mendhekar, S.; Gadge, R.; Kanaskar, S. IoT Based LPG Gas Leakage Detector. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2019**, *5*, 531–534.
14. Santiputri, M.; Tio, M. IoT-based Gas Leak Detection Device. In Proceedings of the 2018 International Conference on Applied Engineering (ICAE), Batam, Indonesia, 3–4 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
15. ALshukri, D.; Sumesh, E.; Krishnan, P. Intelligent border security intrusion detection using iot and embedded systems. In Proceedings of the 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 15–16 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–3.
16. Saquib, Z.; Murari, V.; Bhargav, S.N. BlinDar: An invisible eye for the blind people making life easy for the blind with Internet of Things (IoT). In Proceedings of the 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 19–20 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 71–75.
17. Soh, Z.H.C.; Husa, M.A.A.H.; Abdullah, S.A.C.; Shafie, M.A. Smart waste collection monitoring and alert system via IoT. In Proceedings of the 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Sabah, Malaysia, 27–28 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 50–54.
18. Paul, S.; Banerjee, S.; Biswas, S. Smart Garbage Monitoring Using IoT. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1181–1185. [CrossRef]
19. Association, A.P. Emotion—APA Dictionary of Psychology. Available online: https://dictionary.apa.org/emotion (accessed on 1 July 2021).
20. Tao, J.; Tan, T. Affective Computing: A Review. In *International Conference on Affective Computing and Intelligent Interaction*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 981–995. [CrossRef]
21. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997. [CrossRef]
22. Song, Y.; Dixon, S.; Pearce, M.T.; Halpern, A.R. Perceived and Induced Emotion Responses to Popular Music: Categorical and Dimensional Models. *Music Percept. Interdiscip. J.* **2016**, *33*, 472–492. [CrossRef]
23. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
24. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
25. Zentner, M.; Grandjean, D.; Scherer, K. Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement. *Emotion* **2008**, *8*, 494–521. [CrossRef]
26. Gomez, P.; Danuser, B. Affective and physiological responses to environmental noises and music. *Int. J. Psychophysiol.* **2004**, *53*, 91–103. [CrossRef]
27. Gingras, B.; Marin, M.M.; Fitch, W.T. Beyond Intensity: Spectral Features Effectively Predict Music-Induced Subjective Arousal. *Q. J. Exp. Psychol.* **2014**, *67*, 1428–1446. [CrossRef]
28. Egermann, H.; Fernando, N.; Chuen, L.; McAdams, S. Music induces universal emotion-related psychophysiological responses: Comparing Canadian listeners to Congolese Pygmies. *Front. Psychol.* **2015**, *5*, 1341. [CrossRef]
29. Wanlu, Y.; Makita, K.; Nakao, T.; Kanayama, N.; Machizawa, M.; Sasaoka, T.; Sugata, A.; Kobayashi, R.; Ryosuke, H.; Yamawaki, S.; et al. Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E). *Behav. Res. Methods* **2018**, *50*, 1415–1429. [CrossRef]
30. Fan, J.; Thorogood, M.; Pasquier, P. Emo-soundscapes: A dataset for soundscape emotion recognition. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 196–201.
31. Griffiths, D.; Cunningham, S.; Weinel, J. A self-report study that gauges perceived and induced emotion with music. In Proceedings of the 2015 Internet Technologies and Applications (ITA), Wrexham, UK, 8–11 September 2015; pp. 239–244.
32. Constantin, F.A.; Drăgulin, S. Few Perspectives and Applications of Music Induced Emotion. In Proceedings of the 2019 5th Experiment International Conference (exp.at'19), Funchal, Portugal, 12–14 June 2019; pp. 481–485. [CrossRef]

33. Liu, M.; Chen, H.; Li, Y.; Zhang, F. Emotional Tone-Based Audio Continuous Emotion Recognition. In *MultiMedia Modeling*; He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 470–480.
34. Ooi, C.S.; Seng, K.P.; Ang, L.M.; Chew, L.W. A new approach of audio emotion recognition. *Expert Syst. Appl.* **2014**, *41*, 5858–5869. [CrossRef]
35. Sezgin, M.C.; Günsel, B.; Kurt, G.K. A novel perceptual feature set for audio emotion recognition. In Proceedings of the Face and Gesture 2011, Santa Barbara, CA, USA, 21–25 March 2011; pp. 780–785. [CrossRef]
36. Yang, Y.H.; Lin, Y.C.; Su, Y.F.; Chen, H. A Regression Approach to Music Emotion Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 448–457. [CrossRef]
37. Yang, Y.H.; Chen, H. Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 762–774. [CrossRef]
38. Eerola, T.; Lartillot, O.; Toiviainen, P. Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In Proceedings of the 10th International Society for Music Information Retrieval Conference, Kobe, Japan, 26–30 October 2009; pp. 621–626. [CrossRef]
39. Seo, Y.S.; Huh, J.H. Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications. *Electronics* **2019**, *8*, 164. [CrossRef]
40. Liu, T.; Han, L.; Ma, L.; Guo, D. Audio-based deep music emotion recognition. In *AIP Conference Proceedings*; AIP Publishing LLC: Melville, NY, USA, 2018; Volume 1967, p. 040021, [CrossRef]
41. Soleymani, M.; Caro, M.N.; Schmidt, E.M.; Sha, C.Y.; Yang, Y.H. 1000 Songs for Emotional Analysis of Music. In Proceedings of the ACM International Workshop on Crowdsourcing for Multimedia, Association for Computing Machinery, Barcelona, Spain, 22 October 2013; pp. 1–6.
42. Fan, J.; Tatar, K.; Thorogood, M.; Pasquier, P. Ranking-Based Emotion Recognition for Experimental Music. In Proceedings of the International Society for Music Information Retrieval Conference, Suzhou, China, 23–27 October 2017.
43. Schafer, R. *The Soundscape: Our Sonic Environment and the Tuning of the World*; Inner Traditions/Bear: Rochester, VT, USA, 1993.
44. Schuller, B.; Hantke, S.; Weninger, F.; Han, W.; Zhang, Z.; Narayanan, S. Automatic recognition of emotion evoked by general sound events. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 341–344.
45. Drossos, K.; Kotsakis, R.; Kalliris, G.; Floros, A. Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal. In Proceedings of the IISA 2013, Piraeus, Greece, 10–12 July 2013; pp. 1–6.
46. Bradley, M.M.; Lang, P.J. *The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective Ratings of Sounds and Instruction Manual*; Technical report B-3; University of Florida: Gainesville, FL, USA, 2007.
47. Mathieu, B.; Essid, S.; Fillon, T.; Prado, J.; Richard, G. YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, The Netherlands, 9–13 August 2010; pp. 441–446.
48. Sundaram, S.; Schleicher, R. Towards evaluation of example-based audio retrieval system using affective dimensions. In Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, Singapore, 19–23 July 2010; pp. 573–577. [CrossRef]
49. Fan, J.; Tung, F.; Li, W.; Pasquier, P. Soundscape emotion recognition via deep learning. In Proceedings of the Sound and Music Computing, Limassol, Cyprus, 4–7 July 2018.
50. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135. [CrossRef]
51. Ntalampiras, S.; Potamitis, I. Emotion Prediction of Sound Events Based on Transfer Learning. In *Engineering Applications of Neural Networks*; Boracchi, G., Iliadis, L., Jayne, C., Likas, A., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 303–313. [CrossRef]
52. Ntalampiras, S. Emotional quantification of soundscapes by learning between samples. *Multimed. Tools Appl.* **2020**, *79*, 30387–30395. [CrossRef]
53. Cunningham, S.; Ridley, H.; Weinel, J.; Picking, R. Audio Emotion Recognition Using Machine Learning to Support Sound Design. In Proceedings of the AM'19: 14th International Audio Mostly Conference: A Journey in Sound, Nottingham, UK, 18–20 September 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 116–123, [CrossRef]
54. Cunningham, S.; Ridley, H.; Weinel, J.; Picking, R. Supervised machine learning for audio emotion recognition. *Pers. Ubiquitous Comput.* **2020**, *25*, 637–650. [CrossRef]
55. Drossos, K.; Floros, A.; Giannakoulopoulos, A. BEADS: A dataset of Binaural Emotionally Annotated Digital Sounds. In Proceedings of the IISA 2014, the 5th International Conference on Information, Intelligence, Systems and Applications, Chania, Greece, 7–9 July 2014; pp. 158–163.
56. Drossos, K.; Floros, A.; Giannakoulopoulos, A.; Kanellopoulos, N. Investigating the Impact of Sound Angular Position on the Listener Affective State. *IEEE Trans. Affect. Comput.* **2015**, *6*, 27–42. [CrossRef]
57. Asutay, E.; Västfjäll, D.; Tajadura-Jiménez, A.; Genell, A.; Bergman, P.; Kleiner, M. Emoacoustics: A Study of the Psychoacoustical and Psychological Dimensions of Emotional Sound Design. *J. Audio Eng. Soc.* **2012**, *60*, 21–28.

58. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
59. Lartillot, O.; Toiviainen, P.; Eerola, T. A Matlab Toolbox for Music Information Retrieval. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 261–268. [CrossRef]
60. Lange, E.; Frieler, K. Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automatized Features. *Music Percept.* **2018**, *36*, 217–242.
61. Spiess, A.; Neumeyer, N. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacol.* **2010**, *10*, 6. [CrossRef]
62. Abri, F.; Gutiérrez, L.F.; Siami Namin, A.; Sears, D.R.W.; Jones, K.S. Predicting Emotions Perceived from Sounds. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 2057–2064. [CrossRef]
63. Altman, N.; Krzywinski, M. Points of Significance: Ensemble methods: Bagging and random forests. *Nat. Methods* **2017**, *14*, 933–934. [CrossRef]
64. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422.
65. Siami Namin, A.; Hewett, R.; Jones, K.S.; Pogrund, R. Sonifying Internet Security Threats. In Proceedings of the CHI EA '16: CHI Conference Extended Abstracts on Human Factors in Compting Systems, San Jose, CA, USA, 7–12 May 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 2306–2313, [CrossRef]
66. Datta, P.; Siami Namin, A.; Jones, K.; Hewett, R. Warning users about cyber threats through sounds. *SN Appl. Sci.* **2021**, *3*, 714, [CrossRef]