Int. J. Human–Computer Studies 191 (2024) 103340

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Musician-AI partnership mediated by emotionally-aware smart musical instruments

Luca Turchet [a,*], Domenico Stefani [a], Johan Pauwels [b]

[a] Department of Information Engineering and Computer Science, University of Trento, Italy
[b] Centre for Digital Music, Queen Mary University of London, United Kingdom

ARTICLE INFO

Keywords:
Music information retrieval
Music emotion recognition
Smart musical instruments
Transfer learning
Context-aware computing
Trustworthy AI

ABSTRACT

The integration of emotion recognition capabilities within musical instruments can spur the emergence of novel art formats and services for musicians. This paper proposes the concept of emotionally-aware smart musical instruments, a class of musical devices embedding an artificial intelligence agent able to recognize the emotion contained in the musical signal. This spurs the emergence of novel services for musicians. Two prototypes of emotionally-aware smart piano and smart electric guitar were created, which embedded a recognition method for happiness, sadness, relaxation, aggressiveness and combination thereof. A user study, conducted with eleven pianists and eleven electric guitarists, revealed the strengths and limitations of the developed technology. On average musicians appreciated the proposed concept, who found its value in various musical activities. Most of participants tended to justify the system with respect to erroneous or partially erroneous classifications of the emotions they expressed, reporting to understand the reasons why a given output was produced. Some participants even seemed to trust more the system than their own judgments. Conversely, other participants requested to improve the accuracy, reliability and explainability of the system in order to achieve a higher degree of partnership with it. Our results suggest that, while desirable, perfect prediction of the intended emotion is not an absolute requirement for music emotion recognition to be useful in the construction of smart musical instruments.

1. Introduction

The field of New Interfaces for Musical Expression (NIME) investigates how to progress the design, development, and evaluation of musical interfaces, along with reflective practices on their use (Jensenius and Lyons, 2017; Bown et al., 2009). Digital musical instruments (DMIs) are a central pillar within this domain Miranda and Wanderley (2006). Since electronics made inroads into the art and science of musical instrument making, several DMIs have been invented in both academia and industry, along with applications based on them (Bovermann et al., 2017). One of the research frontiers in the NIME field is represented by the so-called Smart Musical Instruments (SMIs) (Turchet, 2019). This is an emerging class of DMIs characterized by sensors, actuators, embedded intelligence (*i.e.*, artificial intelligence methods running on embedded devices), and wireless connectivity to local networks and the Internet. These self-contained, connectivity, interactive, and intelligent features may confer DMIs with unprecedented context-awareness or proactivity capabilities.

To date, only a handful of musical instruments encompassing some of the envisioned features of SMIs exist in both industry and academy. In particular, research about how to confer intelligence to musical instruments is still in its infancy. Methods to progress the design of SMIs are provided by the recent machine learning developments in the field of Music Information Retrieval (MIR), which investigates computational methods to extract information from musical signals. However, thus far the field of NIME has had a relatively scarce interaction with the field of MIR. As a result, current SMIs have only marginally exploited the possibilities offered by the underlying hardware, and advanced context-aware applications have yet to be created.

In context-aware computing one of the aspects defining the context surrounding a smart object is the type of activity the user is performing with it (Abowd et al., 1999). Concerning musical activities, playing a musical instrument with a given emotional expression is one of the relevant aspects of context that can be detected and subsequently repurposed for proactive applications. Music Emotion Recognition (MER) is a subfield of MIR that deals with the retrieval of emotions contained in musical signals (Yang et al., 2018; Gómez-Cañón et al., 2021; He and Ferguson, 2022). Nevertheless, most of existing MER methods

* Corresponding author.
E-mail addresses: luca.turchet@unitn.it (L. Turchet), domenico.stefani@unitn.it (D. Stefani), j.pauwels@qmul.ac.uk (J. Pauwels).

https://doi.org/10.1016/j.ijhcs.2024.103340
Received 27 March 2024; Received in revised form 25 May 2024; Accepted 16 July 2024
Available online 23 July 2024

have been developed for music involving multiple musical instruments, rather than individual instruments. Directly reusing, for this scenario, existing machine learning models would therefore be unlikely to produce satisfactory results. Moreover, most of the utilized datasets have been created from annotations of listeners rather than from the annotations of the musician who created and performed the music, ideally with varying degrees of the emotional intensity (from weak to strong). Therefore, there is a need for novel datasets specifically created for the task of retrieving emotions from musical excerpts played by musicians on individual instruments. These aspects are crucial for enhancing DMIs with emotional awareness. It is worth noticing that the emotion expressed by a musician may be different from the emotion they actually felt during the act of playing (Van Zijl and Sloboda, 2011). Nevertheless, from the acoustic signal alone it is possible to investigate only the emotion intended by the musician (physiological signals would be a much better indicator for felt emotions) (Turchet et al., 2024).

In this study we aim to create a method for recognizing a musician's intended emotions from the signal of an individual musical instrument, as well as embed such intelligence into the instrument with the end goal of turning it into an emotionally-aware SMI that enables novel application scenarios. This research endeavour is underpinned by the vision for a technology that can forge an effective partnership between a musician and an artificial agent, where the human and the machine can cooperate through the mediation of a musical device. How to design a system able to achieve a partnership truly combining the human and machine complementary skills and capabilities is an unsolved problem for the musical domain, and this is particularly true for the case of emotions in music and the deployment of the artificial agent into the musical instrument itself. A complementary open question concerns the extent to which musicians accept and take advantage of such novel technology, which calls for proper assessment procedures. In summary, our investigation was driven by the following research questions:

RQ1: How well can state-of-the-art MER systems running on an embedded device identify a musician's intended emotion from the signal of an individual instrument?

RQ2: How effective is transfer learning in neural networks for a single-instrument MER task when the donor corpus contains multiple instruments?

RQ3: What is the experience of musicians interacting with an emotionally-aware SMI?

RQ4: Do musicians change their perception of the interaction with an emotionally-aware SMI if the set of classifiable emotions is known in advance?

This work builds upon our previous study reported in Turchet and Pauwels (2022), which investigated the classification of emotions musically expressed by classical and steel-string guitar players. First, we extend that study to investigate two new musical instruments: piano and electric guitar. Second, for the first time we embed the developed MER method within the instrument itself, leading to emotionally-aware smart piano and smart electric guitar. Third, we investigate the actual use of the developed technology by musicians to assess their experience as well as the level of the resulting human-artificial partnership.[1] The intended target user of the developed MER method is a musician playing an SMI. In particular, we focused on the figure of the composer–performer rather than the composer or the performer, because they are the type of musician who commonly creates and expresses emotionally connotated music, such as improvisations (*e.g.*, for recreational music making, performances, or rehearsals). Furthermore, merging the roles of composer and performer removes any possibility of emotional ambiguity between composition and performance, such as when a performer interprets a composition in a way that contradicts the composer's intent.

We selected a transfer learning approach strongly inspired by a state-of-the-art MER model reported in Alonso-Jiménez et al. (2020). It relies on a source model trained for music tagging on a much larger collection (Pons Puig et al., 2018), which is freely available. We implemented our model to run on an embedded system. Before adopting a transfer learning workflow we attempted other MER methods, but without achieving good performance. Our hypothesis was that the use of a transfer learning MER model coupled with a relatively small ad hoc dataset of individual instruments involving four emotions (aggressiveness, relaxation, happiness, sadness) would have led to satisfactory recognition accuracy, thus enabling the creation of emotionally-aware SMIs.

Equipping SMIs with the ability to be aware of the emotions expressed by the player enables the creation of dedicated applications and services supporting various kinds of musical activities. For instance, in the context of Internet of Musical Things applications (Turchet et al., 2018), having reliable machine learning models for emotion recognition would enable the control of peripherals external to the instrument (*e.g.*, stage lights, smoke machines, visuals, haptic devices for the audience), whose evolution will vary depending on the emotion musically expressed by the player. This enables the exploration of new frontiers for multi-sensory music composition and to conceive new concert experiences. Moreover, new applications based on the emotional indexing of large music catalogs could be devised, such as the retrieval of emotional music through SMIs interactions with the cloud. Furthermore, the SMIs' capability of retrieving the emotion contained in the music signal can be exploited to inform in real-time the behaviour of AI agents for music generators, fostering the creation of new ecosystems of humans and machines that enable new forms of artistic expression. In general, the integration of emotion recognition capabilities within musical instruments can spur the emergence of novel art formats and new services for musicians, which calls for further artistic research on such unexplored but promising possibilities.

2. Related work

2.1. Smart musical instruments and embedded audio

While different SMIs products and prototypes have been created by the industry and academic research (noticeable examples are the Sensus Smart Guitar by Elk,[2] the smart acoustic guitar by HyVibe,[3] and the Lava Me 3 by Lava Music[4]), today the idea of enhancing musical instruments with advanced intelligent capabilities remains a vision more than a reality (Turchet, 2019). The barriers that thus far have hampered the creation of SMIs have been the lack of appropriate hardware and software tools for their development (Renney et al., 2022). Nevertheless, in recent years different embedded platforms dedicated to audio processing tasks have been proposed, such as Bela and the Elk Audio OS. Moreover, such availability of increasingly powerful embedded computers has led many deep learning framework developers to devise software optimized to run trained models in resource-constrained contexts (Stefani et al., 2022). As a result, the use of deep learning on embedded audio devices has become more widespread.

These advances in hardware and software technology concretely enable the creation of SMIs and applications for them. Nevertheless, to the authors' best knowledge the challenge of integrating a MER method into an individual instrument has not been addressed yet. The present study is the first to investigate how to turn a conventional musical instrument into an emotionally-aware one.

[1] A demo video of the system in action is available at https://youtu.be/MiHcn7VHEHA

[2] https://youtu.be/fqzEQnsSIoY?si=JnVy8ndarbN3rQ69

[3] https://www.hyvibeguitar.com/

[4] https://www.lavamusic.com/lava-me-3

2.2. Automatic music emotion recognition

Several studies at the confluence of MIR and music psychology have investigated the relations between emotions and specific musical attributes, such as low-level (*e.g.*, spectral features), perceptual (*e.g.*, articulation), and high-level semantic features (*e.g.*, genre) (Panda et al., 2020a). This has allowed scholars to uncover various associations. For instance, it has been found that sadness and anger are often associated with minor modes, while happiness is frequently related to pieces characterized by major modes (Gabrielsson and Lindström, 2001). Moreover, simple, consonant harmonies are typically associated with happiness, pleasantness, or relaxation, whereas complex, dissonant harmonies with emotions such as excitement, tension, or sadness (Laurier et al., 2010).

Different authors have proposed MER methods, fostered by the fact that emotion is one of the most prominent criteria used by listeners to search music (Inskip et al., 2012). Relevant examples of such methods are reported in Laurier et al. (2007), Yang et al. (2008), Laurier et al. (2010), Aljanaki et al. (2017) and Panda et al. (2020b). Typically MER methods are based on two distinct approaches. The first consists of the classification of a given musical excerpt into one or more emotions, thus becoming a multi-label classification problem with a fixed vocabulary (Chowdhury et al., 2019). The second comprises the regression of a continuous emotional space such as the Arousal–Valence one (Russell, 1980), and subsequently clustering such space to obtain a specific emotion vocabulary (Soleymani et al., 2013). In this paper, we focus on the first approach. As shown by results of existing studies (Yang et al., 2018; Panda et al., 2020b) and the Audio Mood Classification task of the 2007–2020 Music Information Retrieval Evaluation eXchange, state-of-the-art solutions for multi-label classifications are still unable to accurately solve simple problems such as the classification of four or five emotion classes.

An important component of the MER endeavours is represented by datasets of music with emotion annotations. Whereas several datasets have been produced by the MIR community (*e.g.*, Yang et al. (2008), Panda et al. (2020b), Aljanaki et al. (2017), Gómez-Cañón et al. (2022)), these typically do not take into account the true nature of the emotions intended by the composer–performer (including the intensity level of the expressed emotions), nor are they focused on individual instruments. Large and freely available emotionally annotated datasets specific to individual instruments and composers–performers are currently missing in the MIR literature, along with dedicated MER methods for such cases. This is a major limitation that hampers the development of emotionally-aware SMIs.

3. Dataset creation

One of the aims of this research was to develop a novel dataset of musical excerpts specifically conceived for the creation of emotionally-aware SMIs. We focused on composers–performers playing new, unfamiliar pieces on two musical instruments: piano and electric guitar. These instruments were selected not only because they are very widespread, but also because they are able to produce radically different timbres (the former purely acoustic sounds, the latter electrically modified sounds resulting from the application of a variety of audio effects). This enables us to assess the presence of variations in the MER algorithm performances due to the timbre of the instrument. In addition, the dataset was conceived to include examples of target emotions with varying levels of intensity, from weak to strong, which enables the study of more subtle variations in emotion.

The focus of the dataset was on four emotions: aggressiveness, relaxation, happiness, and sadness. On the one hand, these emotions were selected because they have been investigated in several studies on emotional expression in music (Gabrielsson and Juslin, 2003), and because they cover the four quadrants of the two-dimensional Arousal–Valence space (Juslin and Sloboda, 2001). On the other hand, they were chosen because they have been tested in previous machine listening setups (Laurier et al., 2010; Alonso-Jiménez et al., 2020; Turchet and Pauwels, 2022).

Table 1
Number of composed pieces in the created dataset categorized by the instrument, composers' emotional intent and its intensity.

Piano recordings					
Intensity	Aggressive	Relaxed	Happy	Sad	Total
Low	21	18	16	18	73
Medium	27	32	30	29	118
High	26	26	28	27	107
Total	74	76	74	74	298
Electric guitar recordings					
Intensity	Aggressive	Relaxed	Happy	Sad	Total
Low	22	29	26	26	103
Medium	33	33	39	36	141
High	39	32	30	32	133
Total	94	94	95	94	377

3.1. Participants

The dataset was created by 23 expert piano players and 27 expert electric guitar players (all Italian; 2 female, 48 male), aged between 18 and 45 (mean = 30.2, SD = 6.1). They reported having at least 13 years of active music expertise (mean = 18.4, SD = 8.4) and on average started learning to play music at the age of 9. We selected such musicians because they were both able to compose and perform emotional intentions well. Specifically, we aimed to avoid potential differences in the intended emotions that may arise between the two roles of composer and performer (Quinto and Thompson, 2013).

3.2. Procedure

Following the approach reported in our previous study (Turchet and Pauwels, 2022), each musician was asked to compose and record at least 12 short emotional pieces, 3 for each of 4 emotions (aggressiveness, relaxation, happiness, sadness). Each recording was required to have a duration ranging from 20 to 60 s and should have been performed in optimal conditions such as in a recording studio or a silent room, using the internal microphone system embedded in the instrument or external microphones. Pianists were asked to not apply any effect to the audio signal either while recording or in the editing production phase, but to use the original acoustic sound of the instrument. Conversely, electric guitarists were instructed to use as much as they wanted any kind of effect and chain of effects. To the end goal of increasing variety in the dataset, all musicians were asked to create multiple pieces within the same emotion that were distinct from one another (*e.g.*, with a different style, tonality, harmonic progressions, etc.). No further indication was given. Therefore, musicians were left completely free to use their creativity to express the indicated emotions, using various degrees of emotional intent (*e.g.*, very sad music or a little aggressive piece), playing technique, expressive technique, style, genre, harmonic progression, tempo, etc.

Most musicians recorded more than the 12 compositions required. This led to a total of 675 recordings, of which 298 for piano and 377 for electric guitar. Subsequently, musicians were asked to indicate for each piece the level of their emotional intent in expressing that emotion, on a 3-point scale indicating a low, medium, and high intensity. For example, regarding sadness the values composers could choose from were "a little sad", "sad", and "very sad" (analogously for the other emotions). Notably, this request was made after and not before the recording because we wanted to leave the musicians free to express the emotion with the intensity that they felt was most appropriate, without imposing a particular level on them. Musicians were not forced to play pieces with all three emotional levels, they were free to skip one or two levels if wanted. Table 1 provides a description of the dataset in terms of instrument, number of composed pieces categorized by the musicians' emotional intent and their intensity.

4. Machine learning models

The purposely created datasets of piano and electric guitar recordings were used to create two separate, instrument-specific music emotion recognition models. For their final deployment as part of an SMI, the entire datasets are used to train each model, and testing is done by letting users play new compositions. However, in order to determine the optimal parameterization of the models, and to get an idea of their performance, a series of five-fold cross-validation experiments were performed with the datasets.

4.1. Experimental setup

The instrument-specific models were created following a strategy inspired by Alonso-Jiménez et al. (2020), which we also followed for the creation of an acoustic guitar-specific MER model in our previous work (Turchet and Pauwels, 2022). We started from a convolutional neural network that encodes audio-specific knowledge in its architecture, called *MusiCNN* (Pons and Serra, 2019). It processes mel-spectrograms in disjoint 3 s chunks. Its pretrained weights obtained from reproducing tags for large collections of audio were used as a starting point to perform transfer learning with our custom, smaller datasets. We used the *MusiCNN* weights up to the penultimate layer and added a new classifier head of two dense layers on top, which we trained in two stages. First the randomly initialized head is trained until convergence with the pretrained weights frozen. Then the whole model is unfrozen to allow a holistic finetuning step.

The main difference with (Alonso-Jiménez et al., 2020), is that we opted for a multi-class approach, where we aim to predict all four emotions concurrently, whereas they created independent binary classifiers for each emotion. Their training data were multi-instrument, downmixed recordings that are commercially available and a magnitude larger in size. We have previously shown that these models do not generalize well to single-instrument recordings (Turchet and Pauwels, 2022). Our motivation to build a single model for all emotions was to make maximal use of our limited data. We also argue this way allows for exploiting mutual information between emotions.

For each 3 s chunk, the model returns values between 0 and 1 corresponding to each emotion due to its final softmax layer. In our controlled setup, we can rely on the fact that the emotion in the recording will remain constant over its whole duration. We therefore average the values of the same emotion for all chunks before determining which of the averages is the highest, a process known as *soft voting* over time. This improves the robustness of the emotion estimate.

From previous human labelling experiments (Turchet and Pauwels, 2022), we have learned that humans deem emotions in music as ambiguous in more than 35% of cases, meaning that two or more emotions are deemed present to the same extent. This presence was measured on a 7-point Likert scale, whereas the output of our machine learning models are continuous values between 0 and 1. The latter means that it is virtually impossible for multiple emotions to be assigned the same output value.

In order to let our machine learning models produce a similar sense of ambiguity, we did not just use the maximum softmax value of the output layer to identify the emotion present, as would be commonly done for multi-class problems. Although it is known that the absolute output values produced by a softmax are unlikely to be calibrated in a way that they can be interpreted as probabilities (Guo et al., 2017), their relative values are a reasonably effective representation of uncertainty (Pearce et al., 2021). Therefore, we assign an unambiguous emotion to a recording only when the distance between the highest and second highest softmax value exceeds a threshold. Otherwise, the emotion is deemed ambiguous. Following our previous work (Turchet and Pauwels, 2022), we set this threshold to $1/7$. Its value can be modified to tweak the proportion of emotionally ambiguous recordings, but the resulting proportion roughly matched the proportion of human ambiguity measured in an acoustic guitar dataset.

4.2. Results

The datasets were divided into five splits, whereas the recordings of each composer–performer were assigned indivisibly to a single split. This way, the scenario of having new performers play with the SMIs was best simulated. The splits were furthermore balanced to have equal duration of emotions. We tried transfer learning with pretrained weights coming from the Magna-Tag-a-Tune (Law et al., 2009) ($\approx 19k$ tracks) and the Million Song Dataset (Bertin-Mahieux et al., 2011) ($\approx 200k$ tracks). Both gave comparable results, so we continued with the Magna-Tag-a-Tune weights. For comparison, we also tried training models from scratch, which resulted in far worse performing models. This illustrates the challenge of capturing wide-ranging concepts such as emotions with comparatively little data. The Python code to perform the experiments and train the final models is available online.[5]

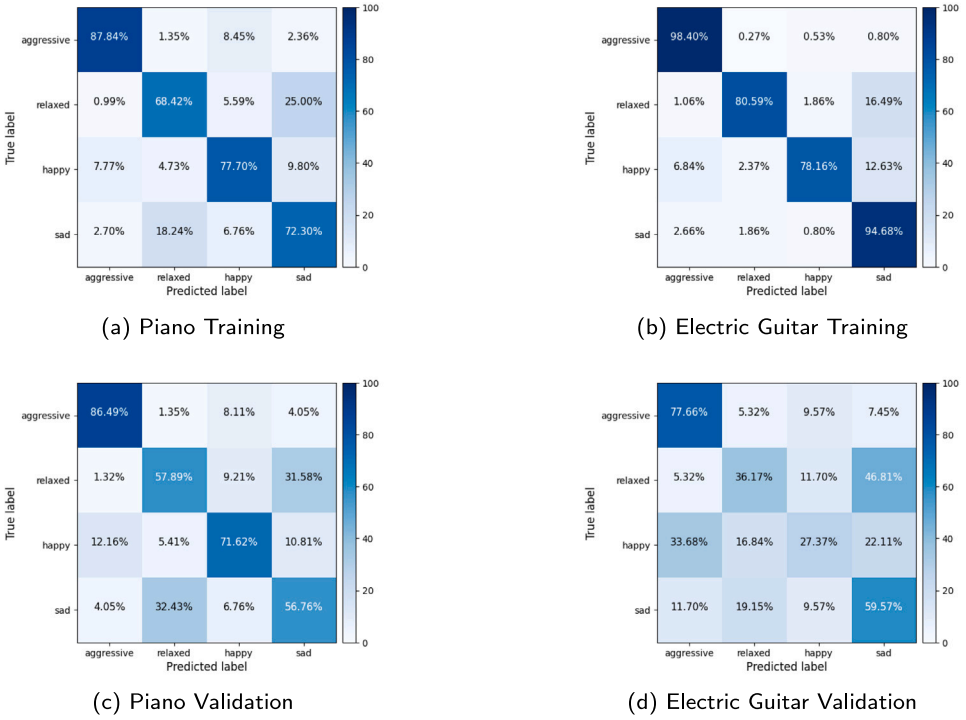
After determining the optimizer, batch size and learning rate hyperparameters, we achieved training accuracies of $76.50 \pm 3.78\%$ and $87.98 \pm 4.47\%$ for piano and electric guitar, respectively. The validation accuracies were $67.95 \pm 3.14\%$ and $50.09 \pm 4.80\%$. All results are reported after the soft voting procedure. The corresponding confusion matrices are depicted in Fig. 1. From these, we observe distinct behaviour between both instruments, despite the identical training procedure. The electric guitar model is clearly overfitting, as shown by the large gap between train and validation accuracies, whereas the piano model behaves much better in this regard. This in spite of the larger number of examples available for electric guitar. It is tempting to ascribe this difference to the wider variety in timbre that an electric guitar augmented with unrestricted effects can obtain, compared to the timbral variety of an acoustic piano (recall the briefing in Section 3.2). However, due to the transfer learning approach, we cannot be entirely sure this is the cause. It might be that the donor weights are simply a more appropriate starting point for one instrument over the other. Nonetheless, the Magna-Tag-a-Tune dataset contains more examples of guitar (though potentially also including acoustic guitar) than piano, and the same goes for our own dataset. It is, therefore, unlikely that a simple imbalance in the amount of training data per instrument is the root cause. An imbalance in training data normalized by data variance (as hard as that is to quantify) is thus more likely.

Regarding differences in accuracy between emotions, we see that aggressiveness is consistently easiest to detect. Again, the caveat needs to be made that this might be due to the pretrained weights, but when we previously noticed this for acoustic guitar emotion recognition (Turchet and Pauwels, 2022), aggressiveness was confirmed by human evaluation to be the most distinct emotion. Relaxed and sadness seem particularly prone to be confused and in the case of electric guitar, a degenerate recognition of happiness around chance level is noted.

The introduction of the ambivalence criterion post-training makes the trends from the confusion matrices more explicit, as seen in Table 2. The easy recognition of aggressiveness leads to low ambivalence values, whereas the other intended emotions are harder to recognize unambiguously. Whether that is true in general or just an artefact of our model cannot be established without large-scale human annotations. The proportion of the ambivalence output differs between instruments, with piano generally having better recognition but with higher ambivalence, and electric guitar worse but more unambivalent emotion recognition.

Regardless of the cause, the introduction of the ambivalence output is intended to improve the user experience of SMI users by giving some leeway to the predictions. We postulate that when the machine learning model is not entirely sure of its output, even though it could be right, it is better to communicate that to the user rather than returning a forced choice that might be completely wrong. This is another example of how uncertainty quantization of machine learning models has real-word applications in music retrieval scenarios, like in Pauwels and Sandler (2019).

[5] https://github.com/jpauwels/instrument_emotion_recognition



(a) Piano Training (b) Electric Guitar Training

(c) Piano Validation (d) Electric Guitar Validation

Fig. 1. Aggregate confusion matrices for the optimal network trained with five-fold cross-validation.

Table 2
Ratio of the machine-predicted emotions as a function of the emotion intended by the composer–performer for both instrument models. Bold indicates a match between predictions and intended emotions. Ambivalent predictions correspond to recordings where the highest machine output did not stand out sufficiently from the outputs of other emotions.

Intention	Piano					Electric Guitar				
	Aggressive	Relaxed	Happy	Sad	Ambivalent	Aggressive	Relaxed	Happy	Sad	Ambivalent
Aggressive	**72.97**	0.00	6.76	0.00	20.27	**75.53**	4.26	8.51	5.32	6.38
Relaxed	1.32	**32.89**	3.95	2.63	59.21	3.19	**31.91**	8.51	36.17	20.21
Happy	10.81	1.35	**50.00**	1.35	36.49	25.26	11.58	**22.11**	18.95	22.11
Sad	4.05	12.16	4.05	**12.16**	67.57	8.51	17.02	7.45	**52.13**	14.89

5. User study

The goal of the user study was to assess the experience of piano and electric guitar players in interacting with the respective emotionally-aware SMI, as well as evaluate the degree of partnership that the technology could achieve. Notably, we investigated the case in which a musician records a short music excerpt and subsequently the MER method embedded in the instrument provides the classification. Our aim was to investigate the musicians' experience of the technology after having created a recording. This is the case occurring when an SMI is equipped with a recording-based service (such as the one reported in Turchet et al. (2020) for content-based queries to online music repositories).

The user study comprised two experiments. In the first experiment, participants were not prompted about what categories of emotions the system could recognize. Conversely, in the second experiment participants were fully aware of the categories. The rationale underlying this choice was to assess whether prior knowledge about the capabilities of the MER system embedded in the instrument could have an impact on the perception of the quality of the system itself and, as a consequence, on the level of satisfaction and acceptance and, ultimately, its usage.

5.1. Apparatus

The emotion recognition system was devised to be executed on a small single-board computer that can be embedded into a musical instrument, turning it into an SMI. We used a Raspberry PI 4 (4 Gb RAM model), which was fitted with a high-resolution analog-to-digital conversion board (Elk PI dev-board) and the Elk Audio OS (Turchet and Fischione, 2021). Elk Audio OS is a Linux-based operating system geared towards low-latency and high-resolution audio processing on embedded platforms. It employs the real-time Xenomai kernel to handle the audio-processing routine of any virtual audio plugin. The design of the whole system easily allows it to be independent of a wired power connection and be embedded into either an SMI or a standalone effect box.

The emotion recognition pipeline is composed of three stages: a recorder, a set of feature extractors, and an instrument-specific deep classifier. The entire pipeline was developed in C++ as a VST audio plugin leveraging the JUCE framework and was cross-compiled for Elk Audio OS and the Linux ARM 64-bit architecture.

The recognition system first records a short emotional piece from the audio input (monaural). Then the recording is downsampled from 48 kHz to 16 kHz and sliced into frames (hop size: 256, frame size: 512). Subsequently, a mel spectrum of 96 bands is computed for each of those frames and disjoint chunks of 187 frames (around 3 s) are created using the Essentia Library (Bogdanov et al., 2013). The type of features used and the parameters for each stage of the extraction are dictated by the architecture of *MusiCNN*, which forms the initial layers of our classifier. The selection of features for the input of *MusiCNN* was informed by studies on musically motivated deep architectures (Pons et al., 2016). In parallel to the extraction, we employ a simple start/stop-of-performance detector based on a signal power threshold and a low-pass filter, to avoid classifying silent sections at the beginning and end of

each recording. Subsequently, three-second-long chunks of the feature matrix are fed one by one to the emotion classifier for the musical instrument played. These models, which were created and trained with TensorFlow and Keras, were converted to the TensorFlow Lite (TFLite) format, and the TFLite interpreter was integrated into the code to execute inference on the embedded device. TFLite was chosen as it was found to be one of the best-performing engines for embedded inference according to the measurements reported in Stefani et al. (2022), and the conversion process from TensorFlow is seamless.

Following the criterion adopted in Turchet and Pauwels (2022), we considered the predicted emotion to be ambivalent whenever two or more of the largest softmax outputs were within 1/7 of each other. For the experiment, the MER system was designed to either produce a single emotion, in the non-ambivalent case, or a list of the emotion labels for which the softmax output was within 1/7 of the largest output.

For the user study, the emotion recognition plugin was controlled remotely via a laptop and a JUCE application through Open Sound Control messages. The remote controller allows an operator to start and stop a plugin for the instrument of choice, set the recording gain level, tune the silence threshold for the start/stop detector, and monitor both a signal meter and the classification results.

5.2. Participants

Eleven expert piano players and eleven expert electric guitar players took part in the experiment (21 Italians, 1 Sri Lankan, 4 females, 18 males, aged between 22 and 59, mean = 32.2, SD = 7.8). They reported having at least 12 years of active music expertise (mean = 16.7, SD = 8.1) and on average started learning to play music at the age of 9. They were selected for their ability to both improvise and perform emotional intentions well. No participant involved during the recordings of the dataset took part in the evaluation experiment. Participants took on average one hour and a half to complete the experiment. The answers provided during the experiments were translated from Italian to English. The procedure, approved by the local ethical committee, was in accordance with the ethical standards of the 1964 Declaration of Helsinki.

5.3. Procedure

The two experiments were conducted in part at the laboratories of the University of Trento, and in part at the houses or recording studios of the participants. Participants were asked to use their own musical instrument and, for the case of electric guitarists, the sound effects equipment they usually utilize (see Fig. 2). During the experiments, participants were assisted by an experimenter for facilitating the unfolding of the procedure.

Experiment 1: unprompted emotions. The first experiment consisted of the following steps.

STEP 1. Participants were briefed about the experiment and signed a consent form. Secondly, they were asked to improvise 4 short musical excerpts, of duration between 20 and 60 s. No indication about what emotions they should play or what emotions would have been recognized by the system was provided. However, they were asked to avoid repeating the same emotion twice, *i.e.*, all 4 recording had to have a different emotional character.

STEP 2. After a piece was recorded, participants were asked to briefly describe the main emotion expressed while playing. They were also asked to rate the intensity of the described emotion on a 3-point scale (low, medium, high).

STEP 3. Subsequently, participants were provided with 2 music excerpts selected among those contained in the dataset reported in Turchet and Pauwels (2022). The aim for this was to test a potential application of the technology where a user performs an emotion-based query-by-playing to an online music repository . Such pieces were intended to match the emotion expressed by the musicians while playing, based on the classification performed by the MER method. The musical excerpts in the dataset were categorized according to the listener annotations in Turchet and Pauwels (2022). In particular, the most frequent annotation for each emotion across multiple listeners (statistical mode) indicated the label emotion to be used. Notably, the resulting label contained more than one emotion whenever most listeners agreed on giving more emotions the same (highest) score. Dataset excerpts were therefore separated into 15 categories, where four identified the main emotions (aggressiveness, relaxation, happiness, sadness) and the remaining 11 all the ambivalent combinations of 2 or more emotions (*e.g.*, aggressiveness + relaxation, aggressiveness + happiness). Each category contained at least 4 excerpts. During the experiment, the emotion, or combination of emotions, predicted by the MER system was used to retrieve 2 tracks from the dataset categories. The arrays of tracks for each category were shuffled at program startup, and a track counter was kept, so that each time, two different random tracks would be played. Notably, the MER system could have been correct or wrong, and even when correct, the retrieved content could have been more or less appropriate depending on the musician's perception. After having listened to both excerpts, participants were asked to rate on an 11-point Likert scale to what extent they were satisfied with the retrieved content (from the sole standpoint of the emotion).

Fig. 2. A picture of a participant of the user study, where it is possible to see the experimental apparatus.

STEP 4. After having completed all recordings and the responses to the questions above, participants were asked to listen to the 4 recordings of their playing. Only at this point participants were exposed to the labels predicted by the classifier, and for each recording they had to rate, on an 11-point Likert scale, to what extent they agreed with the predicted emotion, as well as they were asked to motivate their answer.

Experiment 2: prompted emotions. In the second experiment, conducted immediately after the first, participants underwent the following steps:

STEP 1. Participants were asked to play 8 short musical excerpts, of duration between 20 and 60 s. For each performance, they were prompted to express one of the four emotions in the vocabulary of the classifier, resulting in two excerpts for each. The order was randomized. For each piece, participants were not prompted with a specific intensity, but were asked to indicate it afterward on a 3-point scale (low, medium, high).

STEP 2. After playing each piece, participants were provided with the classification produced by the system and were asked to rate on an 11-point Likert scale to what extent they agreed with it, and why.

Final questions. At the end of both experiments, participants were asked about their demographics as well as the following questions on an 11-point Likert scale [strongly disagree, strongly agree]:

- I enjoyed using this system;
- I would recommend the system to a friend;

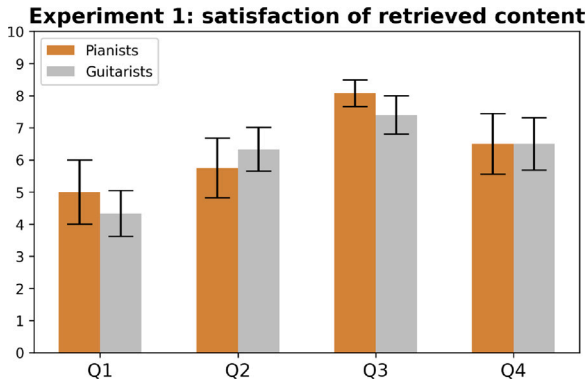


Fig. 3. Mean and standard error of participants' satisfaction with respect to the retrieved content in Experiment 1 (unprompted emotions).

- I would use the system frequently;
- Knowing in advance the range of emotions classifiable by the system impacted positively my experience with the system;
- Knowing in advance the range of emotions classifiable by the system impacted negatively my experience with the system.

The first three questions aimed at assessing the extent to which the system and the concept behind it were appreciated. The last two questions aimed at assessing to what extent the prior knowledge of the system recognition capabilities had an impact on participants' experiences. We asked about both the positive and negative impact to assess the consistency in the participants' answers as well as to avoid biases in the terminology of the question.

The following open-ended questions were also asked:

- Please elaborate on whether the prior knowledge of the emotions classifiable by the system impacted your perception of it compared to when the emotions were not known.
- In which musical activities would you use the system most?
- How would you improve the system?

Moreover, participants were given the possibility to leave an open comment. Finally, a short semi-structured interview was conducted to gather further insights about the experience of interacting with the system, especially from the standpoint of the perceived partnership level.

5.4. Results

5.4.1. Results of experiment 1 (unprompted emotions)

In Experiment 1 participants were not aware of the emotions recognizable the system and could select their own label to describe their intended emotions. Table 3 provides a description of the resulting dataset in terms of number of composed pieces categorized by the composers' emotional intent and their intensity, where the chosen emotional labels were grouped in one of the four quadrants of the Russell's circumplex model of affect (Russell, 1980). Specifically, the following labels were comprised in each quadrant:

Q1: Happy, joy, cheerfulness, vitality, joyful delight, uplifting, optimistic;

Q2: Aggressive, rage, angry, tension, excitement, craving jealousy, afraid;

Q3: Sad, melancholic, nostalgic, distress.

Q4: Relaxed, chill, calm, peaceful, serenity, love;

Table 3
Number of composed pieces in the dataset resulting from Experiment 1, categorized by the (unprompted) composers' emotional intent (across the 4 quadrants of the Russell's model) and its intensity.

Piano recordings					
Intensity	Q1	Q2	Q3	Q4	Total
Low	1	1	1	1	4
Medium	3	7	9	1	20
High	7	4	3	6	20
Total	11	12	13	8	44
Electric guitar recordings					
Intensity	Q1	Q2	Q3	Q4	Total
Low	0	1	1	0	2
Medium	8	6	5	6	25
High	4	5	4	4	17
Total	12	12	10	10	44

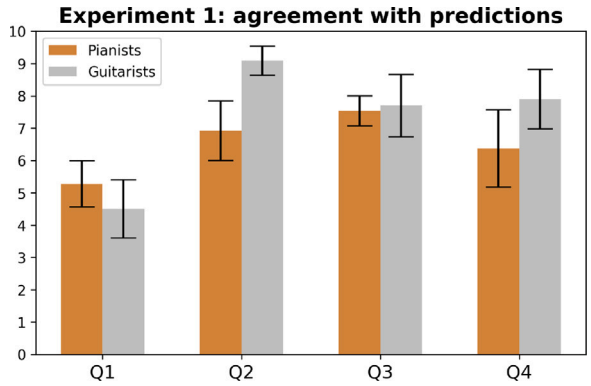


Fig. 4. Mean and standard error of participants' agreement with the predicted labels in Experiment 1 (unprompted emotions).

Table 4 reports the confusion matrix of the labels predicted by the MER method for each of the four categories across which the chosen emotions were grouped. Fig. 3 illustrates the mean and standard error of participants' ratings of the satisfaction with respect to the retrieved audio files. Fig. 4 illustrates the mean and standard error of participants' agreement with the labels predicted by the MER method.

An ANOVA was performed on two different generalized linear mixed effect models, one for the satisfaction ratings and one for the agreement ratings. Specifically, each model had the rating (satisfaction, agreement), quadrant (Q1, Q2, Q3, and Q4), and instrument (piano, electric guitar) as fixed factors, and the playing subject as a random factor. For each model, the assumption on the normality of the residuals was verified. A significant main effect was found only for factor quadrant ($p < 0.001$) not for instrument. A post hoc test, performed on the fitted model using pairwise comparisons adjusted with the Tukey correction, showed that satisfaction ratings were higher for Q3 compared to Q1 ($p < 0.001$) and Q2 ($p < 0.05$). The post hoc test on the agreement ratings showed that these were lower for Q1 compared to Q2 ($p < 0.001$), Q3 ($p < 0.001$) and Q4 ($p < 0.05$).

A further analysis was conducted to assess whether the three level of intensity had an influence on participants satisfaction and agreement ratings. No significant difference was identified.

5.4.2. Results of experiment 2 (prompted emotions)

Table 5 provides a description of the dataset resulting from Experiment 2, in terms of number of composed pieces categorized by the composers' emotional intent (prompted) and their intensity.

Table 6 reports the confusion matrix of the labels predicted by the MER method for each of the four emotions participants were asked to express.

Table 4
Confusion matrix of the labels predicted by the MER method in Experiment 1. Legend: S = sad, H = happy, R = relaxed, A = Aggressive.

		A	H	R	S	SR	SA	SH	HR	HA	RA	RAH	SHR	SHA	SAR	SHRA
Pianists	Q1	5			2	4	1									
	Q2	1	2		3	2							1			2
	Q3				3	5										
	Q4				4	9										
Guitarists	Q1	11				1										
	Q2	10				2										
	Q3	1			1	6	1								1	
	Q4	2			1	5	1								1	

Table 5
Number of composed pieces in the dataset resulting from Experiment 2, categorized by the composers' emotional intent and its intensity.

Piano recordings					
Intensity	Happy	Aggressive	Relaxed	Sad	Total
Low	1	0	1	3	5
Medium	9	7	8	10	34
High	12	15	13	9	49
Total	22	22	22	22	88
Electric guitar recordings					
Intensity	Happy	Aggressive	Relaxed	Sad	Total
Intensity	Aggressive	Relaxed	Happy	Sad	Total
Low	3	2	2	1	8
Medium	12	1	13	12	38
High	7	19	7	9	42
Total	22	22	22	22	88

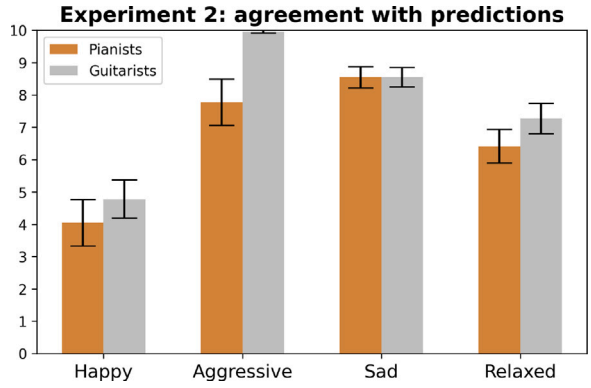


Fig. 5. Mean and standard error of participants' agreement with the predicted labels in Experiment 2 (prompted emotions).

Fig. 5 illustrates the mean and standard error of participants' agreement with the labels predicted by the MER method. An ANOVA was performed on a generalized linear mixed effect model, which had the rating (agreement), emotion (aggressive, happy, relaxed, sad), and instrument (piano, electric guitar) as fixed factors, and the playing subject as a random factor. The assumption on the normality of the residuals was verified. A significant main effect was found only for factor emotion ($p < 0.001$). A post hoc test, performed on the fitted model using pairwise comparisons adjusted with the Tukey correction, showed that the agreements for happy were significantly lower than those for aggressive ($p < 0.001$), relaxed ($p < 0.001$) and sad ($p < 0.001$), as well as those for relaxed were significantly lower than those for aggressive ($p < 0.001$) and sad ($p < 0.01$).

A further in-depth analysis was conducted to assess whether the system accuracy performances varied with the intensity of the expressed emotion, as well as whether such intensity had an influence on participants agreement ratings. No significant difference was identified.

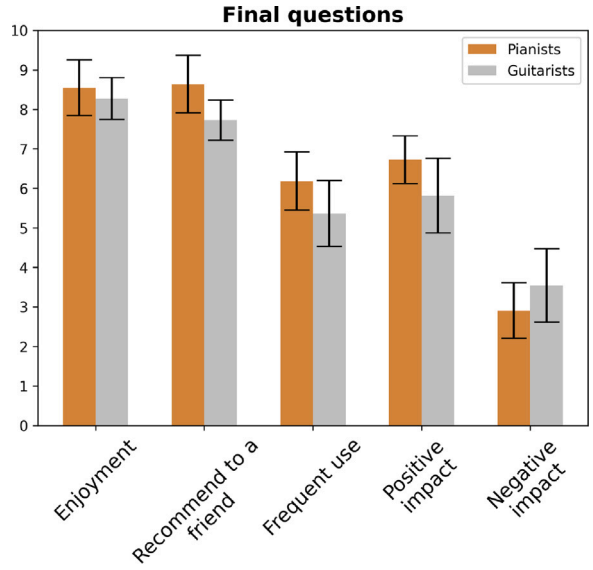


Fig. 6. Mean and standard error of participants' evaluations of the final questions.

5.4.3. Final questions

Fig. 6 illustrates the mean and standard error of participants' evaluations of the questions related to enjoyment, recommendation to a friend, frequency of use, as well as the positive and negative impact that the prior knowledge of the classifiable emotions had on the system experience.

5.4.4. Thematic analysis

The participants' reasons leading to the scores reported in Figs. 4 and 5, as well as the open-ended responses to the final questionnaire items and the short interview, were analysed via a reflexive thematic analysis (Braun and Clarke, 2019). The following themes were identified. Since no significant differences were identified for pianists' themes with respect to the electric guitarists' ones, the thematic analysis was conducted jointly on all participants.

Themes common in Experiment 1 and 2.

Complete satisfaction and total agreement. Of course, all participants expressed their total satisfaction for the system recognition capabilities when the output corresponded to the conveyed emotional intention. Complete agreement with the machine output (score value 10) accounted for 27.58% of the total in Experiment 1 and 34.85% in Experiment 2. In a few cases some participants reported to be in full agreement with the system output even when this did not match their initial intention.

Complete dissatisfaction and total disagreement. For some participants the system output was so far from their expressed intention that reported to be in compete disagreement with the system (*e.g.*, *"I disagree at 100%, in my playing there was nothing sad"*), and wondered why the system produced that output (*e.g.*, *"I do not understand where this judgment comes from"*). The complete disagreement (score value

Table 6
Confusion matrix of the labels predicted by the MER method in Experiment 2. Legend: S = sad, H = happy, R = relaxed, A = Aggressive.

		A	H	R	S	SR	SA	SH	HR	HA	RA	RAH	SHR	SHA	SAR	SHRA
Pianists	A	11	1		4		4								1	1
	H	6	2		5	5							2		1	1
	R				2	20										
	S				3	19										
Guitarists	A	22														
	H	10	1	2		1	1			2					4	1
	R				3	18	1									
	S	1		1	3	14									3	

0) amounted to 5.68% of the total in Experiment 1 and 5.11% in Experiment 2.

Comprehension for system output and partial agreement. By far the most recurrent theme in the participants' comments was that of a partial agreement with the system output. The scores between complete agreement and complete disagreement (score values between 1 and 9) amounted to 66.73% of the total in Experiment 1 and 60.02% in Experiment 2. What is interesting is that the vast majority of participants reported to understand why the system provided such labels, despite the fact they had originally expressed a different kind of emotion. Relevant examples include the following cases: (1) Q1/aggressiveness expressed by playing but sadness or aggressiveness+sadness was returned by the system (*e.g.*, *"My choice of the chords could indeed suggest that the rage originated from sadness"*; *"It was not in my intentions to play also sad but I understand the interpretation of the algorithm"*; *"I recognize that there was a footprint of sadness"*); (2) Q2/happiness expressed by playing but aggressiveness or aggressiveness+happiness was returned by the system (*e.g.*, *"The character of this improvisation was happy, however, I can understand the aggressive classification due to the fast tempo and strong attacks"*; *"I can understand, indeed in some moments I played with energy"*); (3) Q3/relaxation expressed by playing but relaxation+sadness was returned by the system (*e.g.*, *"I can partially agree with the classification of sad, due to the chord progression and the playing style"*; *"It makes sense, there are some passages with elements of sadness"*); (4) Q4/sadness expressed by playing but sadness+relaxation was returned by the system (*e.g.*, *"Yes, in fact I inserted a major chord that could lead to a sense of relaxation"*; *"Correct, it was a relaxed sadness"*; *"The predominant atmosphere was sadness but I acknowledge that there is a part that can be perceived as relaxed"*).

Themes specific to Experiment 1.

Realization after listening. Five participants reported to be in agreement with the system response following the listening of their recordings (*e.g.*, *"I have realized just now while listening to the recording that it was an atmosphere relaxed and then sad, differently from when I was playing where I had intended as just relaxed"*; *"My intention was to be peaceful but upon listening I notice that I have been too aggressive sometimes"*; *"I agree because indeed it was not just sad but also relaxed, I noticed it while I was listening, but when I was playing I intended only sadness"*).

System terminology acceptance. Participants were unaware of the labels returned by the system, but when these were provided to them and differed from those chosen, most of them tended to agree with the system output. This was attributed to the fact that some emotions can be conceived as linked (*e.g.*, *"Melancholy and sadness are emotions liked between each other"*) or compatible (*e.g.*, *"Calm and relaxation are compatible, the difference is subtle"*).

Themes specific to Experiment 2.

System justification and speculations on the reasons. Ten participants reported speculative comments on the reasons why the system produced a particular output, tending to justify the system predictions (*e.g.*, *"The system interprets the music as relaxed because of the slow tempo"*; *"The algorithm may have misinterpreted because of the several staccatos in the chords, the absence of the pedal, and scarce presence of legatos in the melody"*; *"I can go with it: I understand why was also relaxed, the final phase was in major"*; *"It makes sense that there are three emotions, my performance had different parts, so it can be interpreted from more standpoints"*).

System influence on self-judgment. Eight participants reported comments about their playing that gave credit to the system predictions, thus manifesting to be somehow biased by the system responses (*e.g.*, *"It unmasks you! I am basically an 'open book' for it. He has a very high sensitivity"*.; *"I agree with sad, and I understand why it was also deemed relaxed: the system understood a part of the piece that was indeed relaxed. Chapeau! It is like a psychologist!!"*; *"I should have had an approach less 'neurotic'... I maybe had an idea of happiness but then this idea was not well translated.."*.; *"Probably it is my style that in general communicates sadness, I have always a melancholic vein inside"*; *"It is true that it was also aggressive, it was a happy piece played with energy"*).

Themes in the final questions and short interview.

Positive impact of prior knowledge. Fourteen participants reported that the prior knowledge of the system behaviour had a positive impact on their experience of the system compared to when this information was not known (*e.g.*, *"Knowing the range of classifiable emotions facilitated the improvisation as I had an idea what the system would predict while I was playing it"*; *"Knowing the recognizable emotions allows one to compare what he wants to express with the machine result, enabling the establishment of a sort of dialogue with it"*; *"Knowing the classifiable emotions beforehand it was stimulating because it spurred me to express the requested emotion in a focused way"*; *"The prior knowledge led me to reflect in a deeper way on the emotions to be expressed"*).

Negative impact of prior knowledge. Five participants commented on the negative impact of knowing in advance the recognizable emotions (*e.g.*, *"I found it a bit limiting knowing that it could classify those four emotions and any mix of them, as I tried to go for more stereotypical ways of representing certain emotions to make the system to classify it correctly"*; *"I felt more free to express emotions that were not present in the system when I was unaware of the recognizable emotions. However, knowing the emotions beforehand allowed me to understand better how the system reasons"*).

No impact of prior knowledge. Three participants commented that the prior knowledge of the recognizable classes did not lead them to change their interaction with the system (*e.g.*, *"Knowing in advance the emotions had an indifferent impact, I would have done the very same things knowing or not knowing how the system behaved"*).

Concept and usefulness. Sixteen participants consistently provided comments suggesting a strong appreciation for the idea of a musical instrument capable of recognizing the musician's expressed emotions, as well as of the services and application that the system could have in a variety of musical activities. All participants except one commented that they found the system potentially useful for a variety of musical activities. Specifically, four participants envisioned to use the system for searching new sonic content from an online repository for listening, exploring, or practising over the retrieved content (*e.g.*, *"Primarily when listening, to find songs that suit a given emotion"*; *"For practice and expanding performance repertoire"*); nine participants for support to composition activities (*e.g.*, *"I would use it in a music composition/production scenario as a tool to double check whether my music actually meets the intentions meant it to have"*; *"Retrieving music with a given emotion could help me find some inspiration"*); five participants

for pedagogical purposes (*e.g.*, *"A student of composition can have a confirmation about the emotion expressed by the music he composed, or reflect about why the music only in part conveys that emotion"*); three participants for enhanced performances (*e.g.*, *"In a live context where the emotions control multimedia systems such as lights and visuals"*).

Limitations. Nine participants suggested that the system could be improved by allowing for a wider emotions vocabulary (*e.g.*, *"Increasing the number of emotions that can be detected would be nice"*). Two participants recommended to add some personalization mechanisms in order to increase the system performances and as a consequence the user satisfaction (*e.g.*, *"What a musician intends for a given emotion may differ from another musician"*). Nine participants requested to improve the accuracy of the recognition to make the system more reliable (*e.g.*, *"Make sure it recognizes better the playing dynamics (not always a faster rhythm means 'aggressive')"*). Three participants commented on the need of having some kind of explanations from the system about the produced outputs (*e.g.*, *"It would be interesting to have insights on the justification of the output labels considering harmony, melody, rhythm, intensity, etc"*.; *"I would like to have the possibility to see in a more detailed manner how the system classifies the emotions in the various parts of the piece, not just the average"*).

Adaptation to the system. Five participants reported that after the completion of Experiment 2 they somehow understood the behaviour of the system in the recognition process. Being more familiar with the system capabilities they could now adapt to it while playing in order to generate a specific response (*e.g.*, *"After a while I understood how to play so for the system to produce a given output"*; *"After knowing the limitation of the system, I'm more able to trick the system to respond with the labels I want, but it kind of limited the range of playing style I would like to use. For instance, I have to avoid using high gain tone if I want to express any emotions that are not aggressive"*).

Awareness of the system presence. Three participants commented that playing an emotionally-aware SMI made them be aware that a listener was always present (*e.g.*, *"I can't but help ignoring that there is a sensitive ear that is listening to me, but at the same time that's highly stimulating for me"*).

6. Discussion

Firstly, our implementation provides evidence that by leveraging state-of-the-art embedded hardware and software it is possible to create an emotionally-aware SMI, thus providing a successful answer to the research question RQ1. The actual usage of the system during the user study allowed us to answer the research question RQ2: Table 6 confirms what we saw in Table 2, that the instrument-specific emotion recognition models are not entirely successful according to objective metrics. This is not unexpected, as our limited amount of data is unlikely to capture the wide diversity of possible emotional expressions and recording conditions. In particular, for piano it seems to lead to an overprediction of sadness and for electric guitar of aggressiveness. However, the models did not collapse completely, and produced above-chance output, especially when taking the ambivalent output into account.

Most strikingly, users of the SMIs proved to be remarkably tolerant to "objectively incorrect" predictions. No doubt that this is helped by the inherent subjectivity of the concept emotion, and this might be different for other musical concepts like tempo or key (though there still is some subjectivity in those). It does show that, while desirable, perfect prediction of the intended emotion is not an absolute requirement for music emotion recognition to be useful in the construction of SMIs. In addition, higher intensity levels of emotions were not found to be a predictor for better recognition accuracies, a result that is in line with the findings reported in Turchet and Pauwels (2022).

The results of the user study provided insights into the answers to the research questions RQ3 and RQ4 concerning the experience of interacting with the developed technology. In general, no significant differences were found at the quantitative and qualitative level between the responses of pianists and electric guitarists. The vast majority of participants greatly appreciated the idea behind emotionally-aware SMIs and the proposed application of query-by-playing to retrieve emotion-specific music from a repository, as well as found the technology potentially useful for a variety of musical activities. Only one participant was skeptical towards the use of the proposed system and was hesitant in general towards artificial intelligence.

Considering Experiment 1, as shown in Fig. 3 the average ratings of the satisfaction with the audio files retrieved by the system were all above the neutrality, with the exception of Q2 which for both pianists and guitarists was around neutrality. Happiness, which we used as representative for quadrant 2, was the worst recognized emotion for electric guitar in the cross-validation results, which indicates there is some predictive value in the machine learning experiments. A trend similar to that reported in Fig. 3 for the satisfaction of the retrieved content is present in Fig. 4 for the agreement with the predictions.

From a comparison between Figs. 4 and 5 it is possible to notice that participants agreement with the system output followed a trend similar for the cases in which the emotions classifiable by the system were known beforehand or not. This is reflected by the confusion matrices in Tables 4 and 6, that show how the respective recognition accuracies also shared a similar trend. Since both the final model and audio data used for the user study differ from those used for the experiments in Section 4.2, causing covariate shift (Shimodaira, 2000), the increase in ambiguous output can be expected. Nonetheless, trends such as the comparative ease of recognizing aggressiveness are present in both user study and cross-validation experiments.

What emerges from the qualitative analysis on the questionnaire responses and interviews is a shared trend across most participants: that of understanding, totally or in part, the motivations underlying the system responses, even when these were not in line with the emotional intent they wanted to communicate. In the cases in which the predicted emotions did not match perfectly the original emotional intent, most participants speculated about why the system returned a certain output (*e.g.*, relaxation+sadness in place of the intended sole relaxation), essentially providing a justification for it. In some cases participants trusted the system so much to think that it was more correct than their own judgment, thus manifesting to be somehow biased by the system responses. This understanding of the system behaviour led some participants to change their own way of playing in order to make the system return a given prediction (*e.g.*, they avoided certain timbres or playing styles because otherwise the system would have predicted an aggressiveness component in the music).

Concerning RQ4, the prior knowledge of the classifiable emotions had a different impact on musicians. Some deemed that knowing in advance the emotions recognizable by the system had a positive impact on their interaction with it. This was mainly due to the fact that this information allowed them to express a given emotion in a more focused way, especially in order to get from the system the corresponding expected output. Conversely, others felt rather constrained in their expressiveness freedom by the limited amount of recognizable emotions (and combination thereof). Only a few participants reported that the prior knowledge did not have any substantial impact on their interaction with the system.

The proposed system targets musicians with any level of musical expertise, and does not necessitate any particular technical knowledge to be operated. Nevertheless, it is worth mentioning that the actual adoption of the proposed class of musical instruments will depend on different factors. These include the definition of compelling use cases for them, such as new services and new art formats, and their concrete utility perceived by the end users. As with many new instruments developed by the NIME community there is the risk that emotionally-aware SMIs will have difficulties establishing themselves (Morreale and McPherson, 2017). To ensure longevity of such instruments it is necessary to focus on the creation of compelling applications and

services for them, as well as a repertoire and a community of users. Furthermore, it is paramount to ensure trustworthy interactions, adhering to responsible design practices (Piskopani et al., 2023; Brusseau and Turchet, 2024).

Participants indicated different areas of improvement, prominently the need for a larger set of recognizable emotions. Interestingly some participants recommended to equip the instrument with the ability of providing more details about the reasons a certain output was given as well as to produce more correct estimates in order to let them perceive to be interacting with a reliable artificial companion. In particular, some participants suggested to tailor the algorithm for the specific user in order to account for individual differences. In summary, participants indicated that they would be open to build a partnership with the artificial agent, concretely using it as a support for various musical activities, provided that a more accurate and reliable behaviour is made available. Thus, it clearly emerges a request for a more trustworthy artificial partner. First, this characteristic can be achieved by improving the effectiveness of the MER methods, which implies the use of a much larger dataset with a wider pool of musicians, or a dedicated large dataset for a specific musician coupled with the use of finetuning mechanisms. Improvements in model architecture, training or uncertainty quantization would also lead to obvious improvements of the user experience. Second, it can be achieved by adopting explainable AI methods capable of effectively communicating to the user the motivations for a given output. This area is today largely unexplored in the music technology field (Bryan-Kinns et al., 2021), and even less for SMIs (Rossi et al., 2023).

It is worth noticing that our study presents some limitations. Firstly, the dataset has been created involving mostly Italian musicians. Involving musicians from several other countries would likely lead to a more general model devoid of potential biases related to the country of origin of the musicians involved during the training. Moreover, the dataset encompasses recordings made mostly by males. A more general purpose and inclusive model should involve a balanced amount of data from female musicians in order to avoid potential gender-related biases (Holzapfel et al., 2018).

7. Conclusions

The primary objective of the present study was to investigate the concept of emotionally-aware smart musical instruments, a class of musical devices equipped with an artificial intelligence agent able to recognize the emotion contained in the musical signal. This ability can spur the emergence of novel services for musicians. For this purpose we deployed on a device to be embedded in a piano and an electric guitar, a MER method based on transfer learning and on an ad-hoc created dataset.

A user study, conducted with pianists and electric guitarists, revealed the strengths and limitations of the developed technology. On average the proposed concept was appreciated by musicians, who found its value in a variety of musical activities where a reliable artificial partner could support their practices related to the expression of emotions in music. Most of participants tended to justify the system with respect to erroneous or partially erroneous classifications of the emotions they expressed, reporting to understand the reasons why a given output was produced. Some participants even seemed to trust more the system than their own judgments. Conversely, other participants requested to improve the accuracy, reliability and explainability of the system in order to achieve a higher degree of partnership with it.

Several avenues are possible for future work. Firstly, in this study we did not investigate the case in which the MER method is called every few seconds to provide a real-time inference of the expressed emotion. The developed methods could be investigated for real-time scenarios, for instance in Internet of Musical Things settings to enable the control of external peripherals via the musical expression of emotions. This can find application during performance or education activities. Secondly, we plan to improve the machine learning model, potentially based on (Zhang et al., 2023) and uncertainty quantization. Ideally, explainable AI methods would be developed, along with personalization methods specifically tailored for a given user. This has the potential to significantly improve the trustworthiness of users towards the artificial agent, especially with respect to the reasons a given classification is produced and to account for individual differences. Thirdly, we plan to extend the results of this study by involving different types of musical instruments, as well as achieving a gender balance in the dataset creation and user evaluation.

The authors hope that the present study can inspire other practitioners in investigating how to enhance musical instruments with intelligent features and dedicated services based on them so that the field of SMIs can flourish.

CRediT authorship contribution statement

Luca Turchet: Writing – review & editing, Writing – original draft, Validation, Software, Project administration, Methodology, Data curation, Conceptualization. **Domenico Stefani:** Validation, Software. **Johan Pauwels:** Writing – original draft, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abowd, G., Dey, A., Brown, P., Davies, N., Smith, M., Steggles, P., 1999. Towards a better understanding of context and context-awareness. In: International Symposium on Handheld and Ubiquitous Computing. Springer, pp. 304–307.

Aljanaki, A., Yang, Y., Soleymani, M., 2017. Developing a benchmark for emotional analysis of music. PLoS One 12 (3), e0173392.

Alonso-Jiménez, P., Bogdanov, D., Pons, J., Serra, X., 2020. Tensorflow audio models in essentia. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 266–270.

Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P., 2011. The million song dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval. ISMIR.

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Herrera Boyer, P., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J., Serra, X., 2013. Essentia: An audio analysis library for music information retrieval. In: Proceedings of the International Society for Music Information Retrieval Conference. pp. 493–498.

Bovermann, T., de Campo, A., Egermann, H., Hardjowirogo, S., Weinzierl, S. (Eds.), 2017. Musical Instruments in the 21st Century. Springer.

Bown, O., Eldridge, A., McCormack, J., 2009. Understanding interaction in contemporary digital music: From instruments to behavioural objects. Organis. Sound 14 (2), 188–196.

Braun, V., Clarke, V., 2019. Reflecting on reflexive thematic analysis. Qualitat. Res Sport Exerc. Health 11 (4), 589–597.

Brusseau, J., Turchet, L., 2024. Ethics framework for internet musical things. IEEE Trans. Technol. Soc..

Bryan-Kinns, N., Banar, B., Ford, C., Reed, C.N., Zhang, Y., Colton, S., Armitage, J., 2021. Exploring XAI for the arts: explaining latent space in generative music. In: Proceedings of the 1st Workshop on EXplainable AI Approaches for Debugging and Diagnosis.

Chowdhury, S., Vall, A., Haunschmid, V., Widmer, G., 2019. Towards explainable music emotion recognition: The route via mid-level features. In: Proceedings of the International Society for Music Information Retrieval Conference. pp. 237–243.

Gabrielsson, A., Juslin, P.N., 2003. Emotional expression in music. In: Davidson, R.J., Goldsmith, H.H., Scherer, K.R. (Eds.), Handbook of Affective Sciences. Oxford University Press, pp. 503–534.

Gabrielsson, A., Lindström, E., 2001. The influence of musical structure on emotional expression.. In: Juslin, P.N., Sloboda, J.A. (Eds.), Music and Emotion: Theory and Research. Oxford University Press, pp. 223–248. http://dx.doi.org/10.1093/oso/9780192631886.003.0010.

Gómez-Cañón, J.S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., Gómez, E., 2021. Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Process. Mag. 38 (6), 106–114.

Gómez-Cañón, J.S., Gutiérrez-Páez, N., Porcaro, L., Porter, A., Cano, E., Herrera, P., Gkiokas, A., Santos, P., Hernández-Leo, D., Karreman, C., et al., 2022. TROMPA-MER: an open dataset for personalized music emotion recognition. J. Intell. Inf. Syst..

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the 34 Th International Conference on Machine Learning. Sydney, Australia, pp. 1321–1330.

He, N., Ferguson, S., 2022. Music emotion recognition based on segment-level two-stage learning. Int. J. Multimedia Inf. Retriev. 11 (3), 383–394.

Holzapfel, A., Sturm, B., Coeckelbergh, M., 2018. Ethical dimensions of music information retrieval technology. Trans. Int. Soc. Music Inf. Retrieval 1 (1), 44–55.

Inskip, C., Macfarlane, A., Rafferty, P., 2012. Towards the disintermediation of creative music search: analysing queries to determine important facets. Int. J. Dig. Libraries 12 (2), 137–147.

Jensenius, A., Lyons, M., 2017. A NIME Reader: Fifteen Years of New Interfaces for Musical Expression. Springer.

Juslin, P.N., Sloboda, J.A., 2001. Music and emotion: Theory and research. 20, (3), Oxford University Press, viii, 487,

Laurier, C., Herrera, P., Mandel, M., Ellis, D., 2007. Audio music mood classification using support vector machine. MIREX Task Audio Mood Classif. 2–4.

Laurier, C., Meyers, O., Serrà, J., Blech, M., Herrera, P., Serra, X., 2010. Indexing music by mood: design and integration of an automatic content-based annotator. Multimedia Tools Appl. 48 (1), 161–184.

Law, E., West, K., Mandel, M., Bay, M., Downie, J.S., 2009. Evaluation of algorithms using games: the case of music annotation. In: Proceedings of the 10th International Conference on Music Information Retrieval. ISMIR.

Miranda, E., Wanderley, M., 2006. New digital musical instruments: control and interaction beyond the keyboard, vol. 21, AR Editions, Inc.

Morreale, F., McPherson, A., 2017. Design for longevity: Ongoing use of instruments from NIME 2010-14. In: Proceedings of the Conference on New Interfaces for Musical Expression. NIME.

Panda, R., Malheiro, R.M., Paiva, R.P., 2020a. Audio features for music emotion recognition: a survey. IEEE Trans. Affect. Comput..

Panda, R., Malheiro, R., Paiva, R., 2020b. Novel audio features for music emotion recognition. IEEE Trans. Affect. Comput. 11 (4), 614–626.

Pauwels, J., Sandler, M.B., 2019. A web-based system for suggesting new practice material to music learners based on chord content. In: Joint Proceedings of the ACM IUI 2019 Workshops. URL http://ceur-ws.org/Vol-2327/IUI19WS-MILC-1.pdf.

Pearce, T., Brintrup, A., Zhu, J., 2021. Understanding softmax confidence and uncertainty. http://dx.doi.org/10.48550/arXiv.2106.04972, arXiv:2106.04972.

Piskopani, A.M., Chamberlain, A., Ten Holter, C., 2023. Responsible AI and the arts: The ethical and legal implications of ai in the arts and creative industries. In: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. pp. 1–5.

Pons, J., Lidy, T., Serra, X., 2016. Experimenting with musically motivated convolutional neural networks. In: 2016 14th International Workshop on Content-Based Multimedia Indexing. CBMI, pp. 1–6. http://dx.doi.org/10.1109/CBMI.2016.7500246.

Pons, J., Serra, X., 2019. Musicnn: pre-trained convolutional neural networks for music audio tagging. In: Late-Breaking/Demo Session in International Society for Music Information Retrieval Conference.

Pons Puig, J., Nieto Caballero, O., Prockup, M., Schmidt, E., Ehmann, A., Serra, X., 2018. End-to-end learning for music audio tagging at scale. In: Proceedings of the International Society for Music Information Retrieval Conference. pp. 637–644.

Quinto, L., Thompson, W., 2013. Composers and performers have different capacities to manipulate arousal and valence. Psychomusicol. Music Mind Brain 23 (3), 137–150.

Renney, N., Gaster, B., Mitchell, T., Renney, H., 2022. Studying how digital luthiers choose their tools. In: CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Rossi, M., Iacca, G., Turchet, L., 2023. Explainability and real-time in music information retrieval: Motivations and possible scenarios. In: Proceedings of the 4th International Symposium on the Internet of Sounds. IEEE, pp. 1–9.

Russell, J., 1980. A circumplex model of affect.. J. Personal. Soc. Psychol. 39 (6), 1161–1178.

Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Statist. Plann. Inference 90 (2), 227–244. http://dx.doi.org/10.1016/S0378-3758(00)00115-4, URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Soleymani, M., Caro, M., Schmidt, E., Sha, C., Yang, Y., 2013. 1000 Songs for emotional analysis of music. In: Proceedings of the ACM International Workshop on Crowdsourcing for Multimedia. pp. 1–6.

Stefani, D., Peroni, S., Turchet, L., 2022. A comparison of deep learning inference engines for embedded real-time audio classification. In: Proceedings of the 25-Th Int. Conf. on Digital Audio Effects, Vol. 3. DAFx20in22, pp. 256–263.

Turchet, L., 2019. Smart musical instruments: vision, design principles, and future directions. IEEE Access 7, 8944–8963.

Turchet, L., Fischione, C., 2021. Elk audio OS: an open source operating system for the internet of musical things. ACM Trans. Internet Things 2 (2), 1–18.

Turchet, L., Fischione, C., Essl, G., Keller, D., Barthet, M., 2018. Internet of musical things: Vision and challenges. IEEE Access 6, 61994–62017.

Turchet, L., O'Sullivan, B., Ortner, R., Guger, C., 2024. Emotion recognition of playing musicians from EEG, ECG, and acoustic signals. IEEE Trans. Hum.-Mach. Syst. in press.

Turchet, L., Pauwels, J., 2022. Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 305–316.

Turchet, L., Pauwels, J., Fischione, C., Fazekas, G., 2020. Cloud-smart musical instrument interactions: Querying a large music collection with a smart guitar. ACM Trans. Internet Things 1 (3), 1–29.

Van Zijl, A.G., Sloboda, J., 2011. Performers' experienced emotions in the construction of expressive musical performance: An exploratory investigation. Psychol. Music 39 (2), 196–219.

Yang, X., Dong, Y., Li, J., 2018. Review of data features-based music emotion recognition methods. Multimedia Syst. 24 (4), 365–389.

Yang, Y., Lin, Y., Su, Y., Chen, H., 2008. A regression approach to music emotion recognition. IEEE Trans. Audio Speech Lang. Process. 16 (2), 448–457.

Zhang, L., Yang, X., Zhang, Y., Luo, J., 2023. Dual attention-based multi-scale feature fusion approach for dynamic music emotion recognition. In: Proceedings of the 24th International Society for Music Information Retrieval Conference. ISMIR, Milan, Italy, pp. 207–214. http://dx.doi.org/10.5281/zenodo.10265259.