Explainability and Real-Time in Music Information Retrieval: Motivations and Possible Scenarios

Michele Rossi
Dept. of Information Engineering and Computer Science
University of Trento
Trento, Italy
michele.rossi-2@unitn.it

Giovanni Iacca
Dept. of Information Engineering and Computer Science
University of Trento
Trento, Italy
giovanni.iacca@unitn.it

Luca Turchet
Dept. of Information Engineering and Computer Science
University of Trento
Trento, Italy
luca.turchet@unitn.it

2023 4th International Symposium on the Internet of Sounds | 979-8-3503-8254-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59510.2023.10335217

***Abstract*—In recent years, various explainable artificial intelligence methods have been proposed for different Music Information Retrieval (MIR) applications. However, real-time aspects of explainability and the musical applications that they can enable have been largely overlooked by researchers. In this position paper, we propose a vision for real-time explanation systems in the context of MIR applications. We describe three application scenarios where such systems would be useful, although the same concepts can be applied to other musical cases. In the first scenario, we propose a system that provides explanations for an emotion recognition model in the context of music composition and production. In the second scenario, we investigate the benefits of explaining a music genre recognition system for live performances. In the third scenario, we explore the utility of an explanation system for an application in musical instrument learning. Subsequently, we discuss the main challenges associated with the envisioned real-time explanation systems and use cases, especially in relation to the need for easy-to-grasp explanations and to the possible embedding of such systems into smart musical instruments.**

***Index Terms*—Explainability, music information retrieval, smart musical instruments.**

I. Introduction

The field of Music Information Retrieval (MIR) has largely benefited from the advancements in Deep Learning (DL) in the past two decades. DL has been successfully applied to a large variety of tasks such as music genre classification [1], music transcription [2], music recommendation [3], chord recognition [4], and music emotion recognition [5]. However, as the DL models grow in capacity (i.e., the ability to learn and represent complex patterns) and increment the number of parameters, the interpretation and understanding of their internal mechanisms diminish. The so-called transparent machine learning models, such as linear models, decision trees, and rule-based models, are instead usually very easy to interpret but in many cases they are not sufficient to achieve high-accuracy results, especially when dealing with complex tasks [6]. On the other hand, black-box models can usually achieve very high performances, but as said earlier their internal mechanisms are obscured even for the model's designer. For this reason, the search for model-agnostic mechanisms capable of inspecting the behavior of a DL model is gaining more and more interest, as witnessed by its growing research momentum [7].

The need for an explanation highly depends on the field of application of the DL model [8]. For some safety-critical fields, such as medicine [9], autonomous driving [10], and air-traffic management [11], the trustworthiness of the model is crucial. Even if very accurate when evaluated in their test set, DL models can lead to very unpredictable results when dealing with real-world inputs that are not appropriately represented in the dataset used for their training phase [12]. Therefore, many techniques have been developed for a better understanding of the models' behaviors. Some of them, such as Partial Dependence Plots [13] and Accumulated Local Effects Plots [14], are useful for developing an intuition on the general model interpretation as they belong to the family of global model-agnostic methods. Other methods, such as LIME [15] and Shapley Values [16] are local methods as they explain individual models' predictions.

In MIR applications there is a less pressing need for interpretability and explainability of the models when compared to other safety-critical domains [17]. Nevertheless, there are some applications where implementing such systems may lead to large benefits for both the model designer and the final user. The designer may make use of interpretability to gain insights into the model's internal functioning, verify its reliability, and have a better comprehension of the model's internal mechanisms. Moreover, she may gain insights into the feature representation used as input for the model, and discover which features the model relies more on. This process can lead to the development of more compact models, i.e., with fewer parameters, yielding in turn less power consumption and reduced inference time.

The end user of a MIR application may also largely benefit from the model's explanations, even more so if those explanations are provided in real time. This type of information can make the user better understand the reasons behind a specific output of the model, and provide her with more insights on the model's reasoning. The real-time aspect of the system could also enable the user to verify the explanations of the system while changing the input values and adapting her behavior accordingly. To better clarify this concept with

an example, let us consider the case of a musician who is executing a performance with her musical instrument. A model that outputs the emotion associated with her execution could inform the musician about the reason for its response. The user can then modify her playing accordingly and verify if her adjustments impact the model response and why, hence creating a feedback loop between the user and the AI model.

However, thus far real-time aspects of explainability and the applications that they can enable have been largely overlooked by researchers, with only a handful of examples existing in the literature. This is especially true for musical contexts where musicians generate the input for an explainable artificial intelligence (XAI) system [18], [19]. To bridge this gap, this paper proposes a vision for the field of real-time explainability in MIR applications, focusing on the interaction of the musician with the XAI system.

The remainder of the paper is organized as follows. In Section II the main literature contributions are reviewed. Section III explores the principal elements that compose the proposed system. In Section IV we elaborate on three possible concrete application scenarios, and Section V highlights the main issues and challenges. Finally, Section VI summarizes the main contribution of this paper and proposes future research directions.

II. Related work

In the last few years, real-time explainability has been explored for different domains, such as tabular data and images [20], multi-agent systems [21], and networking (specifically, graph neural networks) [22]. However, to the best of our knowledge, only a handful of studies have focused on real-time explanation in the field of Music AI [18], [19]. In fact, a significant body of research emphasizes *offline* approaches, and this is particularly prominent within the domain of MIR.

Choi et al. [23] proposed an explanation mechanism based on auralization. They converted the learned convolutional features obtained from deconvolution [24] into audio signals to make them listenable. They found that the initial layers extract simple audio concepts, such as the onset detector, while the subsequent layers focus on more high-level audio features, such as harmonic and rhythmic textures.

An interesting approach for explaining music emotion recognition with a two-step approach is proposed in [25]. A convolutional model is used to extract perceptually relevant mid-level features that are then fed into a linear model (which is intrinsically interpretable) for the emotions' prediction. The procedure incremented the model explainability at the cost of a small loss in performance compared to an end-to-end approach. For the mid-level features, the authors refer to the dataset proposed in [26].

In [27], the authors proposed an explanation mechanism for classifying playing techniques, such as vibrato, tremolo, and trill. Their approach is based on convolutional networks and layer-wise relevant propagation. They found how the relevant regions of the explanations were associated with the modulation rate of playing techniques, discarding irrelevant features such as the pitch.

Mishra et al. [28] extended the local interpretability method proposed in LIME [15] to the music domain and specifically to the singing voice detection problem. They explored the interpretability with three different input representations: temporal, frequency, and time-frequency. They proved how high-accuracy values do not imply trustable models. They also verified the correspondence between their model-agnostic approach and saliency maps [29].

The explanations associated with time-frequency input representations are usually not easy to interpret. Therefore, the authors of [30] proposed another extension of LIME, where the perturbations were created by adding and subtracting components extracted by a source separation algorithm. This approach enabled the explanation to be listenable.

Some contributions, although not strictly confined to the realm of MIR, have successfully integrated real-time capabilities within the broader domain of Music AI, as exemplified by the two following works. The study reported in [18] explores the importance of feedback between the AI system and the musician during a performance of music improvisations. The authors specifically focused on a collaborative improvising AI drummer that is able to communicate its confidence through an emoticon-based visualization. This research revealed a favorable association between the external communication of the machine's internal state and the level of human engagement in the music performance

The authors of the study reported in [19] delved into the realm of real-time explainability within the generative music domain by expanding upon a prior model known as MeasureVAE [31]. Specifically, their approach enhances model explainability through latent space regularization, enabling specific dimensions to align with meaningful musical attributes. Additionally, they provide a real-time user interface for dimension adjustments and offer visualizations of musical attributes within the latent space to aid users in comprehending and predicting the impact of latent space dimension alterations. Their strategy aims to enhance collaboration between human and AI systems, a field that has gained increasing attention in recent years. This trend is evident in a recent publication [32], where the authors introduced the field of *Explainable Computational Creativity* (XCC) as a subfield XAI. The primary objective of XCC is to establish bidirectional communication channels between humans and Computational Creativity (CC) systems, to promote co-creation and enhance the quality, depth, and utility of their collaborative efforts.

III. Vision

In this section, we propose our general vision of real-time explainability for MIR applications. This is illustrated in Fig. 1. We propose a class of XAI systems that interact with musicians at the moment in which the musical content is generated. The system gets as input the musical signal (symbolic, e.g., MIDI, or in the form of audio) and, at constant temporal intervals, provides the user not only with the result

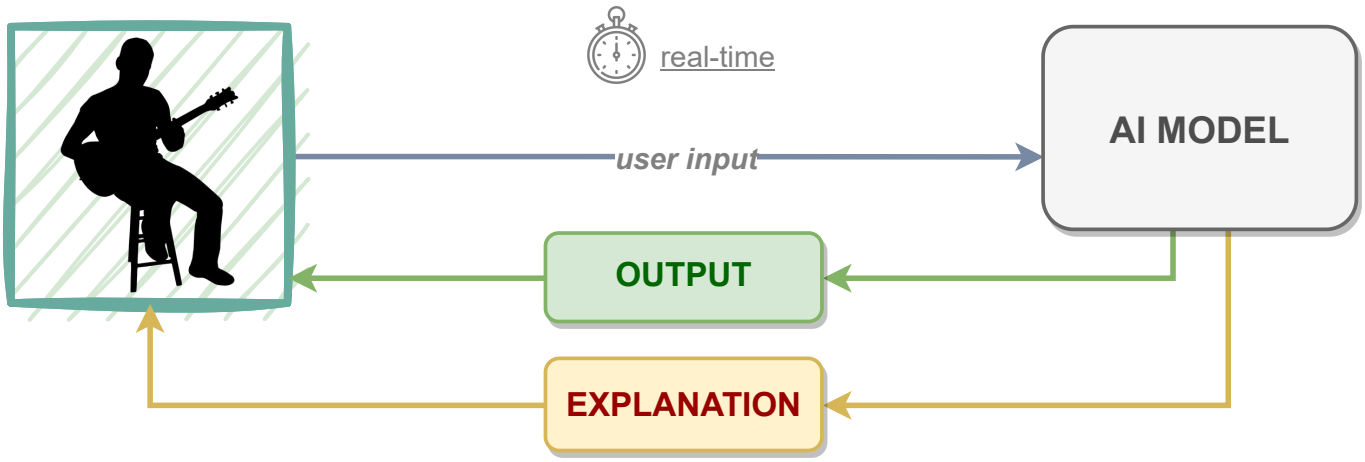

Fig. 1. General structure of a real-time explanation system. The user inputs an audio signal which can be for example an instrumental execution, an entire song, or an instrumental exercise. Then, the system outputs its prediction, which can consist of a label (e.g., genre, mood, and emotion) or an accuracy value. The explanation provides the user with the reason(s) (possibly at different levels, i.e., using low/intermediate/high-level features) associated with the output, which helps the user understand and improve her performance.

of the machine learning model but also with the explanation for such a result. In particular, we contextualize our vision not only in the usage of conventional devices and software (e.g., Digital Audio Workstations running on a laptop) but also in the emerging family of smart musical instruments [33] and the paradigm of the Internet of Musical Things (IoMusT) [34]: the XAI system is directly embedded in the musical instrument itself, which produces the desired output and explanation in real-time and transmits these wirelessly to an external peripheral (e.g., a screen or a smartphone).

As described in Section II, explainability in the context of MIR has been scarcely explored so far in the literature, especially if compared to other domains such as computer vision and natural language processing. However, the reason associated with a specific output of a machine learning model, especially if provided in real-time, can be useful for different categories of users in the music context. Hereinafter, we make use of a practical example to explain the concept more clearly.

Consider for example a music producer who is asked to produce a song with some specific characteristics (e.g., genre, mood, or emotional impact). For the sake of simplicity, let us restrict our analysis to one characteristic: the musical genre. If the producer decides to use an automatic system to verify if the requested criterion is met, in general, she would lack the information (i.e., the reasons) associated with a possible undesired prediction. At this point, she could modify some aspects of the production and reiterate the process, but she would base her modification only on her skills (moreover, this trial-and-error uninformed process would be rather time-consuming and likely ineffective). To improve her work experience, we could consider an implementation similar to the one proposed in [28], which provides an explanation associated with the genre prediction.

However, such an explanation may be difficult to grasp as it could refer to some hardly interpretable audio features. It can be related, for example, to some specific regions of the spectrogram, which is a too low-level piece of information for the producer. In this case, the producer should be provided instead with more meaningful (and higher level) reasons associated with the model output. In other words, explanations need to be appropriate to the specific application (music production in this case) and should provide insights that are useful for the correction of the model's input, if needed. The explanation can be for example associated with specific instruments that cause the model to output a particular prediction, as proposed in [30]. As the authors of this paper describe, this result can be achieved by decomposing the main signal, in our example the song, into meaningful components with a source separation algorithm. Then, an algorithm based on LIME [15] is developed, where the components are utilized as intermediate interpretable features. Therefore, the explanation of the model is interpretable as it refers to components that are meaningful for the user of the application. To clarify this approach even better, let us imagine that the song is divided into five components associated with five possible musical instruments. The explanation of the model could refer, e.g., to the presence of the electric guitar as the main factor that contributes to classifying the genre as rock. Therefore, the producer could know where to intervene if she wants the model prediction to be different.

Now the producer is provided with a genre prediction (e.g., rock) and the reason for that outcome (e.g., the presence of the electric guitar) and can simply intervene on the song according to this information. However, from a practical point of view, there is still an issue with this approach. Every time the producer needs a prediction and the corresponding explanation, the song should be exported and processed by the AI model. Moreover, with this approach, she would receive a single prediction of genre and associated explanation for one song, even if the song consists of many different parts (e.g.,

verse, chorus, and bridge) that can largely differ from one another. Both these issues can be resolved *if the explanation is provided in real-time*. This, in fact, would allow the producer to make modifications *while observing the model outcome and its associated explanation*. By doing so, she can quickly gain information on which actions modify the model's output and the reasons for those changes. Moreover, she could learn which parts of the song contribute to specific musical genres and why. These two aspects can provide useful insights for the user, who could try for instance to learn some input-output relationship discovered by the model. While these relationships may not be so intuitive from a human perspective, the producer could still make use of this acquired knowledge for subsequent productions.

The illustrative example reported in Fig. 1 provides a motivation for the proposed vision. In the next section, some other possible scenarios for the application of real-time explanation are provided, which further justify the need for real-time explainability of MIR methods.

IV. Scenarios

In this section, some practical applications (beyond the example discussed in the previous section) are proposed, contextualized, and discussed.

A. Real-time explainability for emotion recognition in music production

In music, the emotion conveyed by a specific piece of music is a crucial aspect for both the listener and the performer. In recent years, diverse machine learning paradigms have been applied to the task of music emotion recognition [35]–[38], but only a few works explored the related explainability aspect [25], [39]. As far as we know, the real-time aspect of such systems has not been yet explored.

Real-time explainability for music emotion recognition models can be used as a tool for learning. The musician who approaches a specific composition may be interested in knowing if her song or performance reaches the desired goal in terms of emotional impact. Moreover, the apprentice musician may need an explanation associated with the output of the model. This piece of information can in fact give insights into how to correct her composition to obtain the desired emotional impact.

To delve deeper into this approach, let us analyze another example, where a musician wants to compose an original song. Her intent with this composition is to convey a sense of sadness, as this song is inspired by previous personal sad experiences. She may start by writing the lyrics, the main melody, and the chord progression of the new composition. Then, she could choose which instruments are to be included in the recording, and the parts of the song for each of them. Then, as the production advances step by step, she could also be interested to know if her song has the desired emotional impact on the listener. Therefore, she could send the composition to other people to collect feedback on the emotional impact of the song. Moreover, she could ask some more expert friends about the reasons for their specific feedback, especially if their emotion did not match the intended one. Thus, she can get advice on what modifications must be performed in order for the song to transmit more sadness. Some suggestions can be related to the rhythm, the melody, or the choice of the chord, just to give a few examples. The musician can then pick the more convincing suggestions and modify her composition accordingly.

Our proposal is to automate this process of getting valuable real-time feedback on how to modify the song to achieve the desired emotional impact. The composer can make use of an AI-based application that outputs, e.g., every 10 seconds, the label for an input song together with an explanation that should be easy to understand for the user. This goal can be achieved through an explanation based on the work proposed by Chowdhury et al. in [25], where mid-level features, such as dissonance, tonal stability, and rhythmic complexity are used for the explanation of a specific emotion. Therefore, the composer can discover the reason associated with an undesired output of the model. For instance, if the system outputs dissonance as the main element that contributes to the wrong prediction, the artist can restrict her search and concentrate more on the elements that contribute to the dissonance aspect of the composition, such as chord choices. On the other hand, if the model predicts the correct intended emotion, the musician can make use of the explanation to exaggerate even more the song's emotional content in the desired direction. As elaborated in Section III, where we presented the music production example to elaborate the main view of our vision, the real-time aspect of such a system would be crucial. In fact, it would allow the musician to modify and adjust different elements of the song during the listening phase, avoiding the necessity of exporting the entire song to get the model prediction and the associated explanation. Moreover, it would allow the musician to focus on particular parts of the song, and verify their specific emotional impact and explanation.

An interesting expansion of this application would be to include multiple diverse explanations for the model output. Taking advantage of the work proposed by Haunschmid et al. [30], the mid-level interpretable features used for the model's explanation could be directly referred to the specific music instruments, as briefly described in the previous section. Therefore, an emotion prediction of happiness could be directly related to the instrument that conveys that specific emotion in the song. This concept can be expanded to other mid-level features. Exploring which type of mid-level features are most useful for providing explanations for different MIR applications would be an interesting research direction.

B. Real-time explainability for genre recognition during live performance

The performance is an important aspect of the musician's professional life. During the exhibition, musicians are able to transmit energy, convey emotions, and create a specific ambiance. Both for the case of performing an original song or a cover, the artist (or the band) desires a specific impact

on the listener. To obtain this goal, diverse aspects may be considered and analyzed. In this context, we will focus only on one aspect, the musical genre, but a similar approach can be extended to other aspects of the song being performed (such as the mood and the emotional impact).

The proposed real-time explanation system in this context could be used during the band rehearsal. This system would enable the band components to comprehend which adjustments should be made for a specific outcome in terms of musical genre. As before, we propose an example to make the concept clearer. Let us imagine a band that aims to present some classical rock songs and wants to make them more danceable, as may be required by the event where they have to perform. Therefore, they may decide to rearrange some songs to make them less rock and more pop/dance. If they are experts, they may recognize some of the elements that can be modified to adjust the song arrangement to get the desired impact. However, usually, there are different combinations of musical instruments, instrumental parts, and execution intentions that can lead to the same result. Different instruments can have a highly different impact on the final genre and there may exist unexpected combinations of the aforementioned elements which lead to the same genre outcome.

Therefore, for musicians who are not appropriately trained for this task, or for those who want to explore different possibilities to enlarge their horizons, a real-time explanation system for music genre classification can be helpful. As for the case of music emotion recognition, explanations based on human-interpretable mid-level features are required. In this context, reasons based on single instruments can be very useful. In fact, when an instrument is recognized as the most impactful factor for a wrong prediction, different adjustments can be made: for instance, the part of the song where that instrument is played can be modified, and the effects applied to that specific instrument can be added, subtracted, or tuned differently, just to make a few examples.

By analyzing this specific use case, we can also appreciate the importance of the real-time aspect of the proposed system. In fact, as the explanations are provided online, the musicians can try different adjustments and this can be done during the execution of the specific song. This is especially true for those modifications that can be made while performing the song, such as activation/deactivation of effects, different right-hand techniques, and different chord positions, if we consider the electric guitar as an example. The real-time aspect also enables to direct the musicians' attention to specific song segments and concentrate on the parts of the execution that are responsible for the wrong genre prediction.

Similarly to what was mentioned in the music emotion recognition scenario proposed earlier, different human-interpretable features can be provided by the system to improve usability. However, the type of explanation should be chosen according to the scenario considered. In fact, the changes that can be made are different if we consider the context of a band rehearsal compared to the context of music production elaborated earlier. For instance, in the case of the band performance, the number of musicians and instruments is fixed, and the number of effects is often very limited (and applicable only to some instruments).

Another aspect that differentiates this specific scenario from the context of music production is related to the device where such a system is implemented. In fact, during band rehearsals, it may be more convenient to have an external device implementing the system, as it results in a more portable and practical solution. Another extension of this approach would consist of the implementation of one separate device for each musical instrument (internally in the case of smart musical instruments where the intelligence is embedded), which would allow the corresponding musician to receive feedback and explanations specifically related to her execution.

C. Real-time explainability for musical instrument learning

In the third scenario, we consider how a real-time explanation system can be useful in the context of learning a musical instrument. This scenario can be related to MIR as the model should be able to retrieve and elaborate information from a specific execution of the musician, as we will describe later. Being quite an unexplored field, we will first describe why and how machine learning can be used in this context. Then, we will explore the role of the explanations and their utility, especially if provided in real time. The focus of this application is related to the execution of a specific song or exercise with a musical instrument. In order to elaborate on how machine learning can be implemented in this process, we consider the execution of a solo with an electric guitar as an example.

The learning process of a guitar solo usually involves more than one step. Initially, the musician needs to acquire some knowledge related for example to the notes that should be played, the finger positions, and the temporal duration of each note. When the guitar solo (or a specific part of it) is broadly memorized, another critical phase of the learning process begins. At this point, the musician needs to bring the level of her execution from a set of notes being played in sequence to a real performance. This involves many different aspects of the execution to be improved. Some examples are the timing aspect, the cleaning of the execution (avoid undesired noise or notes), the precision in notes' changes, and the correct loudness of each note which is influenced by the right-hand technique. Moreover, there are some more subtle, but very important aspects such as the playing intention, the feeling, and the personal interpretation of the execution. These last aspects are more advanced and less intuitive to grasp, especially for people who are not in the field. From an intuitive level, these improvements have the positive consequence of avoiding a mechanical and cold execution of the guitar solo. In fact, if we imagine an automatic machine that executes a guitar solo playing the correct notes at the exact timing and without dynamic micro-variations, we would immediately recognize the absence of feeling, emotion, intention, and personal interpretation.

For this second step of learning (after the solo is broadly memorized), machine learning methods could be applied to

make an evaluation of the performance execution. In fact, considering a specific guitar solo, the learning process of many guitarists would follow similar paths. While learning, they usually make similar errors related to more concrete and evident aspects in the initial stages and to more subtle and interpretative aspects as the execution improves (as previously described).

Therefore, for the improvement of a specific execution, the guitar teacher finds always similar issues to be solved and proposes appropriate solutions, which are the ones she has found to work best for each specific problem. Machine learning comes into place if we want to automate such a process. This automatic system would be very useful if we consider how most self-taught musicians approach their learning phase. A source of information, which can be a book, a guitar magazine, or, more frequently, an instructional video, is selected and used as a reference. Even if this source can be really valuable, the musician does not receive any feedback on her performance, which instead would be very useful to improve her learning experience and avoid mistakes as the learning phase evolves.

The machine learning system would help the musician by providing a score value related to the performance of the specific solo, together with an explanation of how to improve the performance. There are different methods for the development of a similar application. One approach consists of identifying mid-level features, such as timing and presence of noise, and correct dynamics as intermediate human-interpretable features. Then, following the approach proposed in [25], a two-step Deep Learning model can be trained. Also in this scenario, the real-time aspect can be very helpful. While playing the song, the musician can identify the specific passages that require improvement. Moreover, after each execution, the system could provide a report where, for each segment of the song, the main issues are highlighted. For instance, the system may indicate that the first 5 seconds of the execution have some problems related to the timing aspect, while the subsequent 10 seconds present some undesired noise. This information can be useful for indicating to the musician which parts and aspects of the execution should be improved.

Notably, also in this case the inference and the explainability components could be computed directly inside a smart musical instrument. The result can then be wirelessly sent to an external device, such as a laptop or a smartphone, which visualizes it via a dedicated application.

V. Challenges

The scenarios described in Section IV pose a set of technical and non-technical challenges, which at present prevent the creation of real-time XAI systems for musical applications. In this section, the main challenges of the proposed approach are elaborated. In identifying these challenges we considered the standpoint of both the XAI system designer and the end user. The significance of the latter is closely related to the field of Human-Centered AI, which has gained notable attention and research emphasis in recent years, as evidenced by recent contributions, e.g., in [40].

Challenge 1: Identification of the users' desiderata

To enable end users to understand, trust, and effectively manage their intelligent musical partners, it is first of all necessary to investigate what are the users' desiderata related to a real-time XAI system in musical settings. The identification of the dimensions of end users' explanation needs should be the driving force underlying the design process of the system, especially in terms of what information has to be provided and what form of presentation for this information should be used. This would maximize not just the usefulness of the provided explanation, but also the trust towards the whole AI system. Trust, indeed, is a fundamental factor in musical partnership, especially when the musical activity unfolds in real-time (e.g., performance). It is important to highlight that desiderata may differ for different stakeholders [41], thus it is paramount to conduct investigations for each kind of user of a given real-time XAI system (performer, composer, teacher, student, producer, etc.). On the other hand, it is also important to devise personalized mechanisms accounting for individual differences within the same class of users.

Challenge 2: Explanation representation for MIR

One of the main challenges is related to how the AI-based application provides explanations to the user. Differently from what happens in other domains, such as tabular data, text, or images, the explainability of MIR applications has been scarcely explored so far. The explanations associated with the typical two-dimensional representation of the audio spectrogram are typically useless from a user's point of view. In fact, a specific region of the spectrogram highlighted as responsible for the model's output is very difficult to interpret (especially by musicians lacking a background in signal theory). Therefore, suited mid-level human-interpretable representations should be discovered. As previously described, some solutions in this direction have been proposed, where the explanations are related to perceptually relevant features [25] or to the presence of musical instruments in the song [30].

One aspect that makes it difficult to develop mid-level features is related to the scarce availability of datasets labeled with such features. This is especially true in the proposed case of instrumental learning, where new datasets related to very specific aspects should be developed.

Challenge 3: Explanations for specific applications

Even if reasonable interpretable explanations are discovered for the MIR domain, it should be explored which explanations should be used for each specific application. For example, in the music production context, explanations related to many

different aspects such as the presence of reverb, compression, and equalization, would be beneficial. On the contrary, for band rehearsals, explanations associated with specific instruments could be preferred. Ideally, more levels of explanations should be provided for a single MIR application. This would enable the user to choose the explanation representation(s) more suited for the specific use case. This aspect can be grasped if we consider the case of the artist who decides to produce an original song. Initially, she may need more high-level explanations related to the specific model output, such as which instrument contributes more to the model's prediction. Then, as the production of the song advances, she may be interested in more subtle elements that contribute to the specific outcome, such as the effects applied, the general equalization, and the amount of reverb.

Challenge 4: Application embedding

Another aspect that must be explored is related to the device where the application is implemented. In the context of music production explained earlier, the system can run directly as an application on the computer that hosts the Digital Audio Workstation. In this case, there is no need for portability or implementation in limited-resource devices. On the contrary, for other applications, such as those related to specific musical instruments, it would be convenient to have the system running on a compact portable device. Another possibility consists of integrating the XAI system directly into the musical instrument, as it has been explored recently in the literature on smart musical instruments [33]. The challenging aspects here relate to the need to devise efficient XAI systems for MIR that not only work in real-time but also on embedded devices, which are constrained in terms of computational power and memory. Real-time embedded audio systems running machine learning models is an active area of research (see e.g., [42], [43]), but the development of embedded real-time XAI systems has not been investigated yet to our best knowledge.

Challenge 5: Presentation form

It is crucial to devise effective methods that enable the end user to make sense and take advantage of the explanation generated, while the musical activity unfolds. Notably, providing an explanation in real-time while the musician is playing may lead to an increase in the cognitive load [44]. Therefore, there is a need to conduct research about the best methods that allow one to provide the desired explanation in real time without hampering the ability of the musician to express herself. The definition of effective visualization strategies is fundamental to achieving this goal (e.g., in the form of text, images, or other visual forms). A complementary possible avenue for this quest may concern the use of haptic feedback. Differently from audition and vision, while playing, the sense of touch is mostly an open sensory channel where it may be convenient to provide real-time explanations (as shown for haptic notifications in previous studies [45]). This entails conducting a completely new strand of research about the design, implementation, and evaluation of touch-based methods for explainability purposes.

Another potential source of issue concerns the sensation of feeling somehow judged by the XAI system [46]. This in turn may limit the creativity of the user. Therefore, there is a need to conduct research on human factors as well as the acceptability of such kinds of systems.

Challenge 6: Evaluation methods

In domains other than the musical one, while there is a growing body of literature that has shown the benefits for users of incorporating explanation in AI systems [47], [48], other works have uncovered that there are situations when the added explanations are not always beneficial [49]. This highlights the need for investigations aiming at understanding if and when explanations are necessary or useful. This is especially true for the yet scarcely investigated domain of music. The benefits of the proposed real-time XAI systems for musical stakeholders must be fully investigated. As such, it is crucial to extensively test with the end user the developed XAI methods, especially in the actual context of use rather than in a laboratory setting. For this purpose, novel evaluation methodologies specific to the case of real-time XAI for MIR-based applications must be devised.

VI. Conclusion and Future Work

In this position paper, the possibility of real-time explainability in the context of Music Information Retrieval applications has been explored and analyzed. First, we exposed our vision proposing how to integrate the recent literature discoveries related to audio explainability into practical applications, highlighting the importance of the real-time dimension. Subsequently, we proposed three application scenarios where a real-time explanation system would be beneficial. In the first scenario, we analyzed the impact on the context of emotion recognition applied to the production of an original composition. In the second scenario, we discussed the case of live performances, especially in the context of band rehearsals. In the third scenario, we argued about the positive impact that the envisioned system would have on musical instrument learning, with a specific focus on the guitar. Finally, we identified the main open challenges ahead of us concerning the implementation of the proposed explainable system.

The main future directions we want to elaborate on in this final part are the primary necessary steps that would make our vision realizable. One crucial aspect concerns the realization of new datasets, whose labels could be used as intermediate human-interpretable features, which are especially required if we follow the explainability approach proposed in [39]. This is particularly true in the musical instrument learning application, where labels related to the imprecision of instrumental executions are needed. Data could be collected, e.g., by making use

of the large amount of instrumental executions we can find on online platforms such as YouTube and Instagram. Moreover, a specific main machine learning model should be developed for this scenario (as well as the other ones), as we described in Section IV. Finally, one crucial future work direction is the explanation representation. This aspect will necessarily involve the collaboration of the end user with the machine learning engineer, who should develop explanations that are useful for the specific application. For this purpose, we suggest to adopt a user-centered design approach.

It is the authors' hope that the content of the present study could stimulate discussions about the development of real-time XAI systems for MIR applications, especially in embedded settings.

REFERENCES

[1] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[2] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," *arXiv preprint arXiv:1604.08723*, 2016.

[3] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 627–636.

[4] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *2012 11th International Conference on Machine Learning and Applications*, vol. 2. IEEE, 2012, pp. 357–362.

[5] L. Turchet and J. Pauwels, "Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 305–316, 2021.

[6] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[7] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[8] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.

[9] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[10] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2942–2950.

[11] Y. Xie, N. Pongsakornsathien, A. Gardi, and R. Sabatini, "Explanation of machine-learning solutions in air-traffic management," *Aerospace*, vol. 8, no. 8, p. 224, 2021.

[12] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[13] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[14] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 1059–1086, 2020.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[17] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural computing and applications*, vol. 32, no. 24, pp. 18 069–18 083, 2020.

[18] J. McCormack, T. Gifford, P. Hutchings, M. T. Llano Rodriguez, M. Yee-King, and M. d'Inverno, "In a silent way: Communication between ai and improvising musicians beyond sound," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–11.

[19] N. Bryan-Kinns, B. Banar, C. Ford, C. N. Reed, Y. Zhang, S. Colton, and J. Armitage, "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 (XAI4Debugging@NeurIPS2021)*, 2021.

[20] Y.-N. Chuang, G. Wang, F. Yang, Q. Zhou, P. Tripathi, X. Cai, and X. Hu, "Cortx: Contrastive framework for real-time explanation," *arXiv preprint arXiv:2303.02794*, 2023.

[21] F. Alzetta, P. Giorgini, A. Najjar, M. I. Schumacher, and D. Calvaresi, "In-time explainability in multi-agent systems: Challenges, opportunities, and roadmap," in *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, 2020, pp. 39–53.

[22] D. Pujol Perich, J. R. Suárez-Varela Maciá, S. Xiao, B. Wu, A. Cabellos Aparicio, and P. Barlet Ros, "Netxplain: Real-time explainability of graph neural networks applied to networking," *ITU Journal on future and evolving technologies*, vol. 2, no. 4, pp. 57–66, 2021.

[23] K. Choi, G. Fazekas, and M. Sandler, "Explaining deep convolutional neural networks on music classification," *arXiv preprint arXiv:1607.02444*, 2016.

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[25] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "Towards explainable music emotion recognition: The route via mid-level features," *arXiv preprint arXiv:1907.03572*, 2019.

[26] A. Aljanaki and M. Soleymani, "A data-driven approach to mid-level perceptual musical feature modeling," *arXiv preprint arXiv:1806.04903*, 2018.

[27] C. Wang, V. Lostanlen, and M. Lagrange, "Explainable audio classification of playing techniques with layer-wise relevance propagation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[28] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis." in *ISMIR*, vol. 53, 2017, pp. 537–543.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[30] V. Haunschmid, E. Manilow, and G. Widmer, "audiolime: Listenable explanations using source separation," *arXiv preprint arXiv:2008.00582*, 2020.

[31] A. Pati, A. Lerch, and G. Hadjeres, "Learning to traverse latent spaces for musical score inpainting," *arXiv preprint arXiv:1907.01164*, 2019.

[32] M. T. Llano, M. d'Inverno, M. J. Yee-King, J. Mccormack, A. Ilsar, A. Pease, and S. Colton, "Explainable computational creativity," in *Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)*, 2022.

[33] L. Turchet, "Smart Musical Instruments: vision, design principles, and future directions," *IEEE Access*, vol. 7, pp. 8944–8963, 2019.

[34] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of Musical Things: Vision and Challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.

[35] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[36] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 acm on international conference on multimedia retrieval*, 2018, pp. 135–142.

[37] Y.-H. Yang, C.-C. Liu, and H. H. Chen, "Music emotion classification: A fuzzy approach," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 81–84.

[38] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[39] V. Haunschmid, S. Chowdhury, and G. Widmer, "Two-level explanations in music emotion recognition," *arXiv preprint arXiv:1905.11760*, 2019.

[40] B. Shneiderman, *Human-centered AI*. Oxford University Press, 2022.

[41] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, "What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research," *Artificial Intelligence*, vol. 296, p. 103473, 2021.
[42] D. Stefani, S. Peroni, and L. Turchet, "A comparison of deep learning inference engines for embedded real-time audio classification," in *Proceedings of the Digital Audio Effects Conference*, 2022.
[43] T. Pelinski, R. Diaz, A. L. B. Temprano, and A. McPherson, "Pipeline for recording datasets and running neural networks on the bela embedded hardware platform," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2023.
[44] M. Çorlu, C. Muller, F. Desmet, and M. Leman, "The consequences of additional cognitive load on performing musicians," *Psychology of Music*, vol. 43, no. 4, pp. 495–510, 2015.
[45] L. Turchet and M. Barthet, "Co-design of Musical Haptic Wearables for electronic music performer's communication," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 2, pp. 183–193, 2019.
[46] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios, "What factors contribute to acceptance of artificial intelligence? a systematic review," *Telematics and Informatics*, p. 101925, 2022.
[47] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 7–9.
[48] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, and A. R. Gerlicher, "Explain yourself! transparency for positive ux in autonomous driving," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–12.
[49] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important? a study of deployed, low-cost intelligent interactive systems," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, 2012, pp. 169–178.