# SHOULD WE CONSIDER THE USERS IN CONTEXTUAL MUSIC AUTO-TAGGING MODELS?

Karim M Ibrahim, Elena V Epure, Geoffroy Peeters, Gael Richard

# SHOULD WE CONSIDER THE USERS IN CONTEXTUAL MUSIC AUTO-TAGGING MODELS?

**Karim M. Ibrahim**[1,2] **Elena V. Epure**[2] **Geoffroy Peeters**[1] **Gaël Richard**[1]
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris
[2] Deezer Research
karim.ibrahim@telecom-paris.fr

## ABSTRACT

Music tags are commonly used to describe and categorize music. Various auto-tagging models and datasets have been proposed for the automatic music annotation with tags. However, the past approaches often neglect the fact that many of these tags largely depend on the user, especially the tags related to the context of music listening. In this paper, we address this problem by proposing a user-aware music auto-tagging system and evaluation protocol. Specifically, we use both the audio content and user information extracted from the user listening history to predict contextual tags for a given user/track pair. We propose a new dataset of music tracks annotated with contextual tags per user. We compare our model to the traditional audio-based model and study the influence of user embeddings on the classification quality. Our work shows that explicitly modeling the user listening history into the automatic tagging process could lead to more accurate estimation of contextual tags.

## 1. INTRODUCTION

Tags are a popular way to categorise music in large catalogues in order to facilitate their exploration and music retrieval on demand [17]. Music tags include different categories such as emotions (sad, happy), genres (rock, jazz), instrumentation-related (guitar, vocals), or listening activities (dance, relax, workout). Traditionally, tags were assigned to music items by humans, either through editors or through crowdsourcing. However, with the expanding availability of online music, there have been also increasing efforts towards developing music auto-tagging models, i.e. systems that do not require to manually annotate the tracks [1]. Music auto-taggers are models trained to automatically predict the correct tags for a given music track from the track content. Several models have been proposed that use the audio content, either as raw signal [15, 20, 26] or pre-processed spectrograms [4, 5, 25, 26], to predict the appropriate tags. However, certain tags largely depend on users and their listening preferences, in particular, the tags referring to the context of music listening such as 'running' or 'relaxing' [21]. Thus, traditional auto-tagging models that rely only on the audio content without considering the case where tags depend on users, are not ideal for describing music with user-dependent tags like contexts. Additionally, their evaluation protocol should be also adapted to account for different users.

Previous studies showed that user context has a clear influence on the user's music selection [10, 18]. Hence, context is progressively becoming the focus of music streaming services for reaching a personalized user experience [14]. The user context, e.g. activity or location, can change frequently while listening to music, which leads to changes in user preferences. Consequently, users often need different recommendations. Automatically inferring the user context is often not feasible due to privacy issues. Hence, giving users the option to select a specific context and propose him/her related personalized tracks is a potential alternative [7]. Another use case is automatic continuation or generation of context-specific playlists for each user which are made available to them to select based on their current context [2]. Thus, describing tracks with contextual tags provides a means to improve music exploration and playlist generation in a dynamic way, suitable for the frequent changes in the user context. However, previous work [13] showed that using only the audio might not be sufficient to predict the right contextual tag of a track without putting the user in the loop. Here, we investigate the impact of including user information in context auto-taggers.

In this paper, we propose the following contributions: 1) a dataset of $\sim$182K user/track pairs labelled with 10 of the most common context tags based on the users' contextual preferences presented in Section 2, which we make available for future research; 2) a new evaluation procedure for music tagging which takes into account that tags are subjective, i.e. user-specific in Section 3; 3) an auto-tagging model using both audio content and user information to predict contextual tags in Section 4. Our experiments in Section 5 clearly show the advantage of including the user information in predicting contextual tags compared to traditional audio-only-based auto-tagging models presented.

## 2. DATASET

To properly study the influence of including user information in context auto-tagging models, we need a dataset of

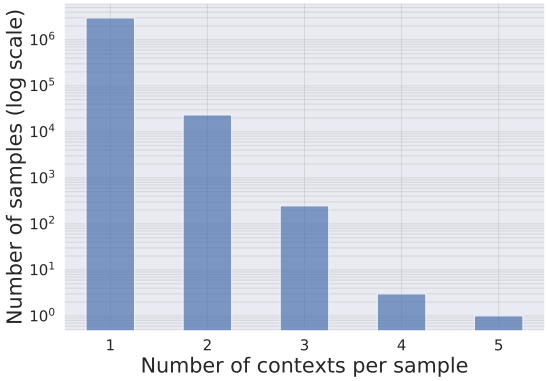


**Figure 1**. Distribution of the number of contextual tags per sample (user/track pair) in the initial dataset.

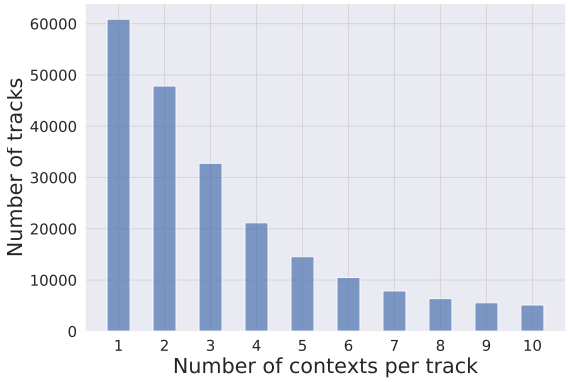


**Figure 2**. Distribution of the number of contextual tags per track in the initial dataset.

tracks labelled with their contextual tags according to different users. For this purpose, we rely on the user-created context-related playlists. Users often create playlists for specific contexts and the titles of these playlists may convey these contexts. Thus, similar to [13,19], we exploit the playlist titles to label tracks with their contextual use. Additionally, we put the users in the loop as playlist creators by explicitly including them in the dataset.

### 2.1 Dataset Collection

To retrieve contextual playlists, we used a set of contextual keywords collected from the literature [9,18,24,27]. Then, we added keywords that were semantically similar. Ninety six keywords were categorized in one of four categories: location, activity, time, and mood. This is similar to the categorization proposed in [14]. To construct our dataset, we selected, out of all collected context-related keywords, 10 which were the most frequent keywords found in the playlist titles in the Deezer catalogue [1]. We selected the keywords that shared a similar number of playlists to avoid any bias due to the popularity of some contexts. The contextual tags we finally selected are: *car*, *gym*, *happy*, *night*, *relax*, *running*, *sad*, *summer*, *work*, *workout*.

We collected all the public user playlists that included any of these 10 keywords in the stemmed title and applied a series of filtering steps to consolidate the dataset, similar to our previous work in [13]. We removed all playlists that contained more than 100 tracks, to ensure that the playlists reflected a careful selection of context-related tracks, and not randomly added. We also removed all playlists where a single artist or album made up more than 25% of all tracks in the playlist, to ensure that the playlist was not intended for a specific artist, similar to [13]. Finally, to properly study the effect of the user on the contextual use of a track, we only kept the tracks that were selected by at least 5 different users in at least 3 different contexts. Hence, our dataset reflects how user preferences change the contextual use of tracks. Finally, we tagged each sample, the track/user pair, with the contextual tag found in the corresponding playlist title.

| | #Samples | #Track | #Users |
|---|---|---|---|
| **Train** | 102K | 15K | 40K |
| **Validation** | 30K | 4.4K | 21K |
| **Test** | 50K | 7.5K | 16K |

**Table 1**. Number of samples (track/user pairs), unique tracks and users in the train, validation and test datasets.

### 2.2 Dataset Analysis

In Figure 1, by observing the distribution of contextual tags per track/user pairs in the dataset, we noticed that most of the pairs were assigned to a unique contextual tag. Let us remind that the log scale is used and a sample represents a user/track pair labelled with the contextual tags. It appears that the majority of users tend to associate a track with a single context. Out of $\sim$3 millions samples, $\sim$2.9 millions are labelled with a single context. Nonetheless, ascertaining if this observation is generally valid requires further empirical investigation. For this study though, we limited our final dataset to track/user pairs with single context tags, i.e. we excluded users that assigned the same track to multiple contexts.

Observing the distribution of contextual tags per tracks in Figure 2, we find that tracks often have multiple contexts associated with them. This shows that the suitability of a track for a specific context varies from user to user. However, as previously outlined, given the user, the track is most frequently associated with the same unique context.

The final dataset for this study contains $\sim$182K samples of user/tracks pairs made of $\sim$28K unique tracks and $\sim$75K unique users. We collected the dataset such that each context is equally represented, ensuring a ratio of $\sim \frac{1}{10}$ of all user/track pairs. We split our dataset in an iterative way to keep the balance between classes across subsets, while preventing any overlap between the users and minimising the overlap between tracks in these subsets [22]. The distribution of our final split dataset is shown in Table 2.2. The dataset is publicly available to the research community [2]

[1] Deezer is an online music streaming service: www.deezer.com

[2] https://doi.org/10.5281/zenodo.3961560

## 3. PROPOSED EVALUATION PROTOCOL

Previous studies on music auto-tagging [4, 20] performed the evaluation in a multi-label classification setup, therefore focusing on assessing the correctness of the tags associated with each track. This is suitable for datasets and tags that are only music-dependent. However, in the case of tags that are also user-dependent, the previous evaluation procedures are limiting.

### 3.1 User Satisfaction-focused Evaluation

The purpose of our study is to measure the influence of leveraging the user information on the quality of the prediction of contextual tags. Consequently, we are interested in measuring the potential satisfaction of each user when predicting contexts, instead of relying on a general evaluation approach that could be biased by highly active users or by the popularity of certain tags. Hence, we propose to compute the model performance by considering each user independently. To assess the satisfaction of each user, the evaluation metrics are computed by considering only the contextual tags specific to a user. Then, to assess the overall user satisfaction, we average the per-user results yielded by each model.

Formally, let $\mathbb{U}$ denote a finite set of users in the test set, $G_u = \{0,1\}^{n_u \times m_u}$ denote the groundtruth matrix for user $u$, $n_u$ is the number of tracks associated with the user $u$, and $m_u$ is the number of contextual tags employed by the user. Similarly, $P_u = \{0,1\}^{n_u \times m_u}$ denotes the matrix outputted by the model for all active tracks and contextual tags for the given user $u$. First, we compute each user-aware metric, hereby denoted by $S$, for a given user $u$ as:

$$S_u = f(G_u, P_u) \tag{1}$$

where $f$ is the evaluation function. In our evaluation, we use standard classification metrics such as the area under the receiver operating characteristic curve (AUC), recall, precision, and f1-score [12]. While the protocol is defined for the general case of multi-label setting, in our current work, given the dataset, it is applied to the case of single-label. Then, we compute the final metrics, by averaging over all users in the test set:

$$S_{\mathbb{U}} = \frac{1}{N} \sum_{u \in \mathbb{U}} S_u, \text{ where } N = |\mathbb{U}|. \tag{2}$$

### 3.2 Multi-label Classification Evaluation

In this work, we develop a system that takes both the audio and the user information as input. As seen in Section 2.2, for a given track and user, there is a single groundtruth context to be predicted. The problem is said to be single-label. However, if we want to compare this system with a system that only takes audio as input, we need to consider during training various possible groundtruth contextual tags for a track, each from a different user. Then, the problem becomes multi-label. The comparison of the two systems is therefore not straightforward. Indeed, for the user-agnostic case, we can train a multi-label system, i.e. a system with a set of sigmoid output activations optimized with a sum of binary cross entropy, and estimate it either as single-label by taking the output with the largest likelihood, or as multi-label by selecting all outputs with a likelihood above a fixed threshold. For these reasons, in the current evaluation, we consider the following scenarios:

1. Multi-output / multi-groundtruth (MO-MG): This is the classical multi-label evaluation where the model outputs several predictions and each track is associated with several groundtruths. This evaluation is however independent of the user.

2. Multi-output / single-groundtruth (MO-SG): In this scenario, a model trained as multi-label (such as a user-agnostic model) is still allowed to output several predictions. However, since the groundtruth is associated with a given user, there is a single groundtruth. The obtained results are then over-optimistic because the model has several chances to obtain the correct groundtruth.

3. Single-output / single-groundtruth (SO-SG): this is the case that is directly comparable to our single-output user-aware auto-tagging model. As opposed to the MO-SG scenario, models trained as multi-label are now forced to output a single prediction, the most likely contextual tag. This prevents them from being over-optimistic as they only have one chance to obtain the correct groundtruth, as does the single-label model too.

## 4. PROPOSED MODEL FOR CONTEXTUAL TAG ESTIMATION

We propose to build a user-aware auto-tagging system. Given that contextual tags are interpreted differently by different users, we hypothesize that considering the user information in training a personalized user-aware contextual auto-tagging model may help. For this, we propose to add to the system, along with the audio input, a user input. We study the effectiveness of representing the user via 'user embeddings', obtained from user listening history.

### 4.1 Traditional Audio-based Auto-tagger

In this paper, we chose the prevalent audio-based auto-tagging model proposed by Choi et al [4]. The model is a multi-layer convolutional neural network. The input to the network is the pre-processed Mel-Spectrogram of the music track. This multi-label classification model predicts, for a given track, the set of all possible tags.

We trained the network with the Mel-spectrogram as an input of size 646 frames x 96 mel bands, which corresponds to the snippet from 30 to 60 seconds for each track. The output is the predictions for the 10 contextual tags. The input Mel-Spectrograms is passed to a batch normalization layer then to 4 pairs of convolutional and max pooling layers. The convolutional layers have a fixed filter size of (3x3) and (32, 64, 128, 256) filters respectively, each followed by a ReLu activation function. The max pooling

filters have a size (2x2) each. The flattened output of the last CNN layer is passed to a fully connected layer with 256 hidden units with ReLU activation function. We apply a dropout with 0.3 ratio for regularization. Finally, we pass the output to the final layer of 10 output units each with a Sigmoid activation function. The loss function is the sum of the binary cross entropy optimized with Adadelta and a learning rate initialized to 0.1 with an exponential decay every 1000 iterations. We applied early stopping after 10 epochs in case of no improvement on the validation set, and kept the model with the best validation loss.

### 4.2 Proposed Audio+User-based Auto-tagger

The Audio+User model that we propose is an extension of the Audio-based auto-tagger described above. Our model has two branches, one for the audio input and one for the user embeddings input. The audio branch is identical to the one described above, i.e. 4 pairs of convolutional and max pooling layers with ReLu activation. The input to the user branch is the user embedding of size 256. We apply batch normalization to it followed by a fully connected layer with 128 units and Relu activation. We concatenate the output of the audio branch and the user branch after applying batch normalization to each. We pass the concatenated output to a fully connected layer with 256 hidden units with ReLu activation function and apply a dropout with 0.3 ratio for regularization. The final layer is made of 10 output units with a Softmax activation function. We train the model with minimizing the categorical cross entropy using the same configuration as in the previous model, described in Section 4.1. We present the flowchart of the complete model in Figure 3

### 4.3 User Embeddings

The user embeddings are computed by applying implicit alternating least squares matrix factorization (ALS) [11, 16] on the users/tracks interactions matrix. The matrix represents the user listening count of the tracks available, with an exponential decay applied based on the lapse since the last listening, i.e. the more recent and frequent a track is listened to, the higher the interaction value. The user embedding is represented as a 256-dimensions vector.

However, the user listening histories are proprietary and represent sensitive data. Additionally, the detailed derivation of the embeddings is an internal procedure at Deezer for the recommendation algorithm. Hence, in order to allow the reproducibility of the current work, we directly release the pre-computed embeddings for the anonymized users present in our dataset.

## 5. RESULTS

We evaluate the two models according to the evaluation protocol proposed in Section 3. First, we evaluate the audio based model with the 3 scenarios: MO-MG, MO-SG, SO-SG. Then, we evaluate the User+Audio model in the SO-SG scenario. Last, we perform the user satisfaction-based evaluation on both models for the SO-SG scenario. In all evaluation protocols, the metrics were macro-averaged.

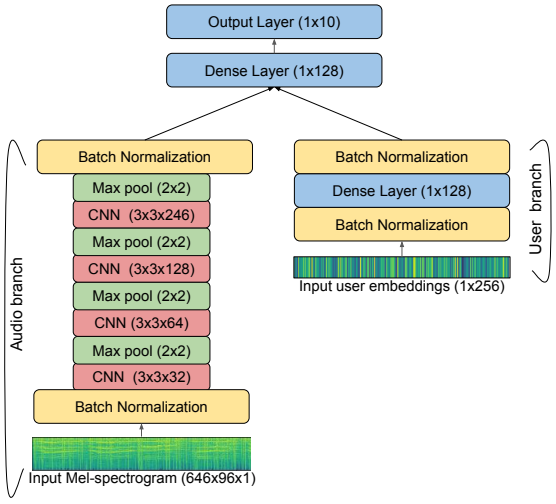


**Figure 3**. Architecture of the Audio+User-based model.

| | AUC | Recall | Precision | f1-score |
|---|---|---|---|---|
| car | 0.56 | 0.96 | 0.47 | 0.63 |
| gym | 0.71 | 0.87 | 0.58 | 0.7 |
| happy | 0.58 | 0.87 | 0.37 | 0.52 |
| night | 0.59 | 0.97 | 0.48 | 0.64 |
| relax | 0.77 | 0.8 | 0.61 | 0.69 |
| running | 0.65 | 0.91 | 0.56 | 0.69 |
| sad | 0.77 | 0.72 | 0.54 | 0.61 |
| summer | 0.6 | 0.97 | 0.61 | 0.75 |
| work | 0.53 | 0.99 | 0.47 | 0.64 |
| workout | 0.75 | 0.84 | 0.52 | 0.64 |
| average | 0.65 | 0.89 | 0.52 | 0.65 |

**Table 2**. Results of the audio-based model (multi-label outputs) on the user-agnostic dataset (multiple groundtruth), MO-MG scenario.

### 5.1 Audio-based Multi-output Multi-groundtruth (MO-MG Scenario)

Table 5.1 shows the results of the audio-based multi-label classification model on our collected dataset without considering the user. The results are consistent with previous studies on context auto-tagging [13]. They show that certain contexts are easier to predict using only the audio input. These are general contexts with similar music style preferences by different users, e.g. 'gym' and 'relax'. By contrast, other contexts are harder to predict from audio only as users listen to more personalized music, e.g. 'work' and 'car'. In consequence, we hypothesis that the variance of the AUC scores across contexts is related to the context dependency on users. Precisely, some contexts could depend more on users than others, making the latter harder to classify without considering the user information.

### 5.2 Audio-based Multi-output Single-groundtruth (MO-SG Scenario)

Table 5.2 shows the results of the same audio-based multi-label classification model which we now evaluate considering the user. The same audio track will now be presented

| | AUC | Recall | Precision | f1-score |
|---|---|---|---|---|
| car | 0.54 | 0.87 | 0.09 | 0.17 |
| gym | 0.66 | 0.6 | 0.18 | 0.27 |
| happy | 0.57 | 0.67 | 0.08 | 0.14 |
| night | 0.57 | 0.6 | 0.11 | 0.19 |
| relax | 0.74 | 0.53 | 0.25 | 0.34 |
| running | 0.6 | 0.57 | 0.15 | 0.23 |
| sad | 0.75 | 0.52 | 0.21 | 0.3 |
| summer | 0.58 | 0.78 | 0.17 | 0.29 |
| work | 0.52 | 0.55 | 0.09 | 0.15 |
| workout | 0.71 | 0.41 | 0.17 | 0.24 |
| average | 0.62 | 0.61 | 0.15 | 0.23 |

**Table 3**. Results of the audio-based model (multi-label outputs) on the user-based dataset (single ground-truth), MO-SG scenario

| | AUC | Recall | Precision | f1-score |
|---|---|---|---|---|
| car | 0.54 | 0 | 0.03 | 0.001 |
| gym | 0.66 | 0.44 | 0.17 | 0.24 |
| happy | 0.57 | 0 | 0 | 0 |
| night | 0.57 | 0.004 | 0.14 | 0.007 |
| relax | 0.74 | 0.6 | 0.23 | 0.33 |
| running | 0.6 | 0.05 | 0.15 | 0.07 |
| sad | 0.75 | 0.003 | 0.16 | 0.006 |
| summer | 0.58 | 0.36 | 0.18 | 0.25 |
| work | 0.52 | 0 | 0.2 | 0 |
| workout | 0.71 | 0.13 | 0.18 | 0.15 |
| average | 0.62 | 0.16 | 0.14 | 0.11 |

**Table 4**. Results of the audio-based model (forced to single-label output) on the user-based dataset (single ground-truth), SO-SG scenario

several times to the system, i.e. for each user who has annotated this track. While the groundtruth is now single-label and will change for each user, the system will output the same estimated tags independently of the user, i.e. the system does not consider the user as input. We observe a sharp decrease in the precision of the model due to false positive predictions for each user. Indeed, since the output of the system is multi-label, it will output several labels for each track, many of them will not correspond to the current user. The high recall of the model shows that it often predicts the right contextual use for many users. However, it also predicts wrong contexts for many other users. That is due to the limitation of the model which predicts all suitable contexts for all users.

### 5.3 Audio-based Single-output Single-groundtruth (SO-SG Scenario)

Table 5.3 shows the results of the same audio-based multi-label classification model when restricted to a single prediction per track. While this is not the real-world case of using the audio-based model, it allows a direct comparison to the single-label User+Audio based model. In this case, we see a sharp drop in the recall due to the limitation of a single prediction per track.

| | AUC | Recall | Precision | f1-score |
|---|---|---|---|---|
| car | 0.61 | 0.12 | 0.13 | 0.13 |
| gym | 0.71 | 0.16 | 0.24 | 0.19 |
| happy | 0.64 | 0.22 | 0.12 | 0.16 |
| night | 0.61 | 0.03 | 0.14 | 0.05 |
| relax | 0.76 | 0.41 | 0.29 | 0.34 |
| running | 0.69 | 0.26 | 0.22 | 0.24 |
| sad | 0.83 | 0.5 | 0.33 | 0.4 |
| summer | 0.65 | 0.2 | 0.3 | 0.24 |
| work | 0.58 | 0.03 | 0.12 | 0.04 |
| workout | 0.75 | 0.37 | 0.2 | 0.26 |
| average | 0.68 | 0.23 | 0.21 | 0.2 |

**Table 5**. Results of the audio+user model (single-label output) on the user-based dataset (single ground-truth), SO-SG scenario.

| | Accuracy | Recall | Precision | f1-score |
|---|---|---|---|---|
| Audio | 0.21 | 0.204 | 0.243 | 0.216 |
| Audio+User | 0.254 | 0.246 | 0.295 | 0.26 |

**Table 6**. Comparison of user-based evaluation for the two models

### 5.4 Audio+User Single-output Single-groundtruth (SO-SG Scenario)

Table 5.4 shows the results for the proposed Audio+User model. Comparing these results with the ones presented in Table 5.3, we observe that the model is performing better than the audio-based model for almost all metrics and labels. The f1-score almost doubles when adding the user information. Additionally, for certain labels as *car, happy, running, sad, summer, work*, the influence of adding the user information is obvious compared to all cases of audio-based evaluation when comparing the AUC values. This is consistent with our hypothesis that for certain labels the influence of user preferences is much stronger than for other labels.

### 5.5 User Satisfaction-focused Scenario

Finally, we assess the user satisfaction by evaluating the performance of the two models on each user independently. We replace the AUC metric with accuracy because AUC is not defined in the case of certain users where a specific label is positive for all samples. Table 5.5 shows the average performance of each model when computed per user. In this case, we observe how the Audio+User model satisfies the users more on average in terms of all evaluation metrics. By investigating the recall and precision, we noticed that our model results in a larger number of true positives, i.e. predicting the correct context for each user, and a lower number of false positives, i.e. less predictions of the wrong contextual tags for each user. The audio-based model is prone to a higher false positives due to predicting the most probable context for a given track regardless of the user. To sum up, including the user information in the model has successfully proven to improve the estimation of the right contextual usages of tracks.

## 6. CONCLUSION AND FUTURE WORK

Predicting the contextual use of music tracks is a challenging problem with multiple factors affecting the user preferences. Through our study, we showed that including the user information, represented as user embeddings based on the listening history, improves the model's capability of predicting the suitable context for a given user and track. This is an important result towards building context-aware recommendation systems that are user-personalized, without requiring the exploitation of extensive user private data such as location tracking [3]. However, there is still large room for improvement to successfully build such systems.

Our current model relies on using the audio content, which is suitable for the cold-start problem of recommending new tracks [6, 23]. However, constructing representative user embeddings requires active users in order to properly infer the listening preferences. Future work could investigate the impact of using different types of user information, such as demographics [8], which could be suitable for the user cold-start or less active users too.

Additionally, we focused on the case of a single contextual tag for each user and track pair. In practice, a user could listen to the same track in multiple contexts, i.e. tag prediction would be modelled a multi-label classification problem at the user level. Future studies could further investigate this more complex case of adding the user information in the multi-label settings.

Finally, while we have proven the advantage of our system on a subset of contexts. Extending the study to a larger number of possible contexts still needs to be addressed. In reality, users listen to music in more diverse contexts, adding levels of complexity to the addressed problem.

## 7. ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

## 8. REFERENCES

[1] Thierry Bertin-Mahieux, Douglas Eck, Francois Maillet, and Paul Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.

[2] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018.

[3] Zhiyong Cheng and Jialie Shen. Just-for-me: An adaptive personalization system for location-aware social music recommendation. In *Proceedings of International Conference on Multimedia Retrieval*, 2014.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

[5] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[6] Szu-Yu Chou, Yi-Hsuan Yang, Jyh-Shing Roger Jang, and Yu-Ching Lin. Addressing cold start for next-song recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

[7] Douglas Eck, Paul Lamere, Thierry Bertin-Mahieux, and Stephen Green. Automatic generation of social tags for music recommendation. In *Advances in neural information processing systems*, 2008.

[8] Bruce Ferwerda and Markus Schedl. Investigating the relationship between diversity in music consumption behavior and cultural dimensions: A cross-country analysis. In *UMAP*, 2016.

[9] Michael Gillhofer and Markus Schedl. Iron maiden while jogging, debussy for dinner? In *International Conference on Multimedia Modeling*. Springer, 2015.

[10] Alinka E Greasley and Alexandra Lamont. Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae*, 15(1):45–71, 2011.

[11] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016.

[12] Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[13] Karim Ibrahim, Jimena Royo-Letelier, Elena Epure, Geoffroy Peeters, and Gael Richard. Audio-based auto-tagging with contextual tags for music. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[14] Marius Kaminskas and Francesco Ricci. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2-3):89–119, 2012.

[15] Taejun Kim, Jongpil Lee, and Juhan Nam. Sample-level cnn architectures for music auto-tagging using raw waveforms. In *Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018.

[16] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[17] Paul Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37:101 – 114, 2008.

[18] Adrian C North and David J Hargreaves. Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition*, 15(1-2):30, 1996.

[19] Martin Pichl, E. Zangerle, and G. Specht. Towards a Context-Aware Music Recommendation Approach: What is Hidden in the Playlist Name? In *Proceddings of International Conference on Data Mining Workshop (ICDMW)*, 2015.

[20] Jordi Pons Puig, Oriol Nieto, Matthew Prockup, Erik M Schmidt, Andreas F Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2018.

[21] Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[22] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2011.

[23] Andreu Vall, Hamid Eghbal-zadeh, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. Music playlist continuation by learning from hand-curated examples and song features: Alleviating the cold-start problem for rare and out-of-set songs. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, 2017.

[24] Xinxi Wang, David Rosenblum, and Ye Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.

[25] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serrc. Data-driven harmonic filters for audio representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[26] Minz Won, Sanghyuk Chun, and Xavier Serra. Toward interpretable music tagging with self-attention. *arXiv preprint arXiv:1906.04972*, 2019.

[27] Karthik Yadati, Cynthia Liem, Martha Larson, and Alan Hanjalic. On the automatic identification of music for common activities. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 2017.