2022 IEEE International Conference on
Systems, Man, and Cybernetics (SMC)

MATT: A Multiple-instance Attention Mechanism for Long-tail Music Genre Classification

Xiaokai Liu*, Shihui Song*, Menghua Zhang§, Yafan Huang†‡
*School of Cyber Science and Engineering
†School of Computer Science and Technology
§ School of Energy and Power Engineering
Huazhong University of Science and Technology, Wuhan, China
{liuxk, songsh, iemhzhang, hyfshishen}@hust.edu.cn

2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-5258-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMC53654.2022.9945477

***Abstract*— Imbalanced music genre classification is a crucial task in the Music Information Retrieval (MIR) field for identifying the long-tail, data-poor genre based on the related music audio segments, which is very prevalent in real-world scenarios. Most of the existing models are designed for class-balanced music datasets, resulting in poor performance in accuracy and generalization when identifying the music genres at the tail of the distribution. Inspired by the success of introducing Multi-instance Learning (MIL) in various classification tasks, we propose a novel mechanism named Multi-instance Attention (MATT)[1] to boost the performance for identifying tail classes. Specifically, we first construct the bag-level datasets by generating the album-artist pair bags. Second, we leverage neural networks to encode the music audio segments. Finally, under the guidance of a multi-instance attention mechanism, the neural network-based models could select the most informative genre to match the given music segment. Comprehensive experimental results on a large-scale music genre benchmark dataset with long-tail distribution demonstrate MATT significantly outperforms other state-of-the-art baselines.**

Keywords: Music Information Retrieval, Music Genre Classification, Long-tail Recommendation

I. Introduction

Music Genre Classification (MGC) [1], which aims to identify the genres of given music segments, is an essential task in the research field of Music Information Retrieval (MIR). Recently, inspired by the progress achieved by Deep Learning (DL) techniques, researchers [2], [3] have widely applied DL techs in the fields of MIR and achieved promising results in various tasks, especially MGC. However, since the neural networks are mostly data-hungry, their performance is heavily subject to the scale and quality of training data. Although such techs achieve very good results on common genres, their performances degrade drastically while extracting long-tail genres, which indicates they routinely suffer from data insufficiency. Moreover, most of the previous works only focus on the benchmarks with class-balanced data. These facts lead the classification of long-tail music genres to be a very challenging problem.

Long-tail music genres cannot be ignored as they contain rich musical information. Besides, the data with long-tail distribution is quit common in real-world settings. As a widely used music research dataset, Free Music Archive (FMA) can be used for evaluating the performance of MGC models. As demonstrated in Figure 1, nearly 82% of the genres in FMA have only a few examples. It means that the MGC models need to be able to identify the genres with the limited number of training instances.

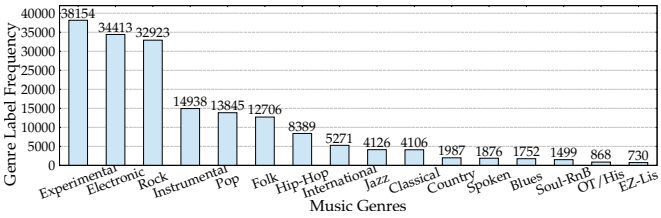


Fig. 1. Music Genre Distribution of FMA Dataset.

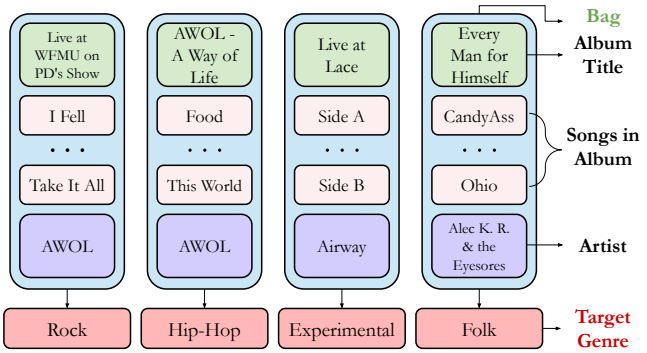


Fig. 2. An Example of MIL for Music Genre Classification.

Inspired by the achievements of multi-instance learning and attention mechanism in various scenarios [4], [5], we propose a novel multiple-instance attention mechanism (MATT) to accurately process those long-tail genres. Similar to the data preprocessing schema in multi-instance learning, we construct the corresponding bag-level dataset using the album-artist pair as the key of the music segment bag. Figure 2 shows the structure of our designed music segment bags for multi-instance learning. The music segments with the same album ID and artist ID would be put into the corresponding album-artist bag since the genres of the segments with the same album ID and artist ID would be the same. By leveraging the bag-level multiple-instance training mechanism, the long-tail problem should be alleviated. Besides, we employ a multiple-instance attention mechanism to help the neural networks identify the genres of the given music segments. With the multiple-instance attention mechanism

‡ Corresponding Author
[1] Github: https://github.com/JohannesLiu/Music-Genre-Classification

utilized in various feature-based and neural network-based models, all of these models achieve better performance. To the best of our knowledge, this is the first work to comprehensively evaluate the performance of various state-of-art music genre classification models on long-tail music genre classification benchmarks.

The key contributions of this work are summarized as follows.

- We propose the multiple-instance learning method for music genre classification, which alleviate the long-tail effect of the music genre distribution. The multiple-instance learning for music genre classification uses the album ID and artist ID pair as the key of the music bag, to build the bag-level dataset. By training models on the built training dataset, the models outperform those without adopting multiple-instance learning.
- We propose the MATT for the long-tail music genre. MATT can calculate the attention scores of each music segment in the bag to help identify the long-tail music genre of the given music segments. With the MATT, the neural networks-based models have better performance than other models without MATT adopt.
- We conduct comprehensive experiments to evaluate the overall performance and long-tail music genre classification performance of the proposed MATT. We also use various metrics to evaluate data-imbalanced classification tasks to compare the performance of MATT with those of the state-of-art feature-based and neural networks-based methods. The results demonstrate MATT achieves state-of-the-art performance.

We arrange the rest of this paper as follows: In section II, we describe the related works about classifying balanced music genres and the long-tail genre classification. In section III, we present our proposed methods in detail. In Section IV, we describe the experiment design and analyze the results. The threads to validity are also included in this section. At last, we conclude this paper and discuss our future works.

II. Related Works

Conventional statistic learning-based models [6]–[8] are hard to meet the requirements of dealing with massive data in production environment. Recently, the deep learning-based music genre classification models [3], [9]–[15] have been widely adopted for MCG and have achieved promising performances. In 2009, [3] built a MGC model using convolutional deep belief network architectures. Then, [9] leveraged the rectifier convolutional neural networks to extract informative features from audio data. [10] used the bidirectional long short term memory (BiLSTM) and recurrent neural networks (RNN) on singing voice detection. The works, which were combined with different structures of neural networks, achieved better performance than the traditional neural networks. [11] leveraged both CNNs and RNNs in their works and introduce convolutional recurrent neural networks (CRNNs). The clustering augmented learning method (CALM), proposed by [12], also achieved promising performance. The CALM is based on the concept of simultaneous heterogeneous clustering and classification to obtain more effective music representations from the audio features extracted via the LSTM autoencoder. [13] proposed an improved technique called CRNN in Time and Frequency dimensions (CRNN-TF), which captures spatial dependencies of music signals in both time and frequency dimensions in multiple directions. Considering the aforementioned limitations, [14] proposed a hybrid architecture, named the parallel recurrent convolutional neural network (PRCNN). [15] proposed bottom-up broadcast neural network which transfers more suitable semantic features for the decision-making layer to discriminate the genre of the unknown music clip. These works mainly focus on the class-balanced genre classification, regardless of the effect of long-tail genres.

However, to the best of our knowledge, there are only a few researches on long-tail MGC tasks [16]–[18]. Choi et al [16] proposed to leverage zero-shot learning to handle unseen labels such as newly added music genres or semantic words that users arbitrarily choose for music retrieval. Urbano et al [17] exploited the combined knowledge, from audio and tagging, using a hybrid representation that extends the track's tag-based representation by adding semantic knowledge extracted from the tags of similar music tracks. Valerio et al [18] proposed a resampling approach to face the class-imbalance problem applied to music genre classification. Although these works have made a positive exploration in the task of long-tail music genre classification, their performance is still not satisfactory. Here we adopt the multiple-instance attention mechanism for music to further improve the performance of MGC models.

III. Methodology

In this section, we first introduce the notations used in this work and briefly explain our workflow. Then we present our multiple-instance attention mechanism-based models for music genre classification step by step.

A. Notations

Inspired by Multiple-instance Learning [19] from the natural language processing field, we split all musical audio segments into multiple album-artist bags and denote them as $\{\mathcal{S}_1, \mathcal{S}_2, \dots\}$. Each bag $\mathcal{S}$ contains multiple instance $\{s_1, s_2, ...\}$ with the same artist ID $\mathcal{P}$ and album ID $\mathcal{A}$. Besides, each instance s in these bags is encoded in mp3 format with a sample rate of 44.1KHz and a sample size of 320 kb/s with stereo channels.

B. Overall Workflow

Our model consists of three major components as shown in Figure 3.

a) Music Instance Preprocessor: Given an instance and its related album ID and artist ID pair, we employ multi-feature extraction methods to embed music audio segments into continuous vector space. In this study, Log-amplitude Mel-spectrogram, MFCC, Chroma, and more than 10 audio feature extraction methods in total are used to extract music audio features.

b) Music Instance Encoder: The music audio encoder is responsible for encoding the low-dimensional music audio vector representation. The Convolutional Recurrent Neural Network is used to implement the music instance encoder for the neural network-based models. For the statistical learning-based model driven by music feature engineering, the music

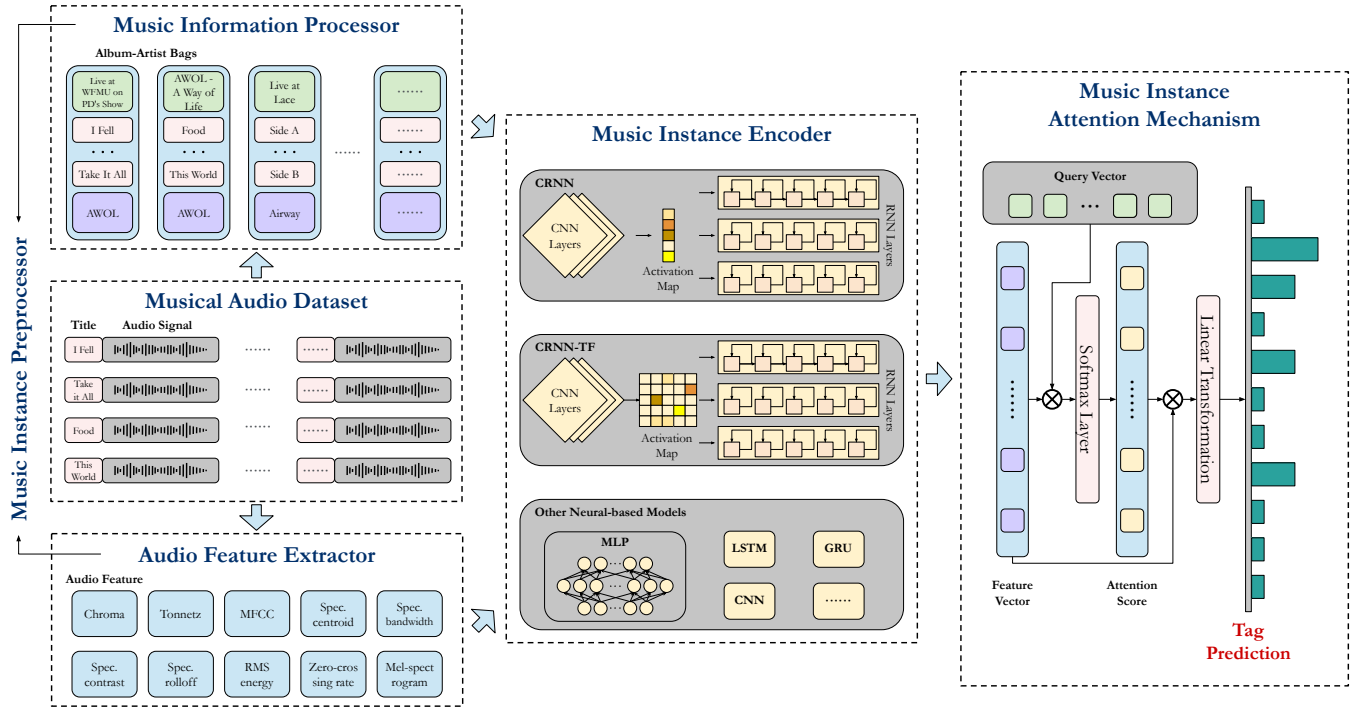


Fig. 3. The Overall Workflow of the MATT-based Music Genre Classification Model.

instance is omitted because the low-dimensional features are not suitable for training.

c) Multi-instance Attention Mechanism: Under the guidance of the final audio embeddings, MATT can identify the most informative music segment exactly matching the relevant genre.

C. Music Instance Preprocessor

a) Music Information Processor: The music information processor aims to process the music dataset and converts them to the corresponding bag level. This process follows the principle of MIL. Specifically, the processor splits all musical audio segments into multiple album-artist bags, where each music segment in the bag shares the same album ID and artist ID. The reason why we leverage MIL to process the music dataset is that the music segments with the same album ID and artist ID always belong to the same genre, which is observed in whole FMA dataset.

b) Audio Feature Extractor: The audio feature extractor is designed to embed music segments into continuous digital representations (i.e. vectors). We extract features from the music audio by 10 methods, which are Chroma, Tonnetz, MFCC, Spec. centroid, Spec. bandwidth, Spec. contrast, Spec. rolloff, RMS energy, Zero-crossing rate feature, and log-amplitude Mel-spectrogram. We implement 9 of the above methods via the Python library – Librosa. As for log-amplitude Mel-spectrogram, its output is a 96×1360 matrix of Mel-spectrogram, where each row and column of this matrix corresponds to a Mel-frequency scale and a Mel-frequency time frame, respectively.

D. Music Instance Encoder

The encoding layer aims to convert given instances into their latent vectors and maintains their semantics at the same time. In this work, we choose neural networks with convolutional layers, e.g. the CRNN and CRNN-TF, to encode input embeddings extracted from the log-amplitude Mel-spectrogram feature extraction method. The reason is that the mel-spectrogram feature has a higher dimension compared with other types of features, which can be smoothly computed by CRNN and CRNN-TF. For other features obtained from the other 9 feature extraction methods, due to the small size of dimensions, they can be processed by both CRNN/CRNN-TF and other encoding methods such as MLP. Other networks such as recurrent neural networks [10] can also serve as audio encoders.

a) CRNN: The convolutional recurrent neural network (CRNN) consists of convolutional layers and Gated recurrent unit (GRU) layers. GRU is a gating mechanism in RNN. The GRU and RNN have very similar structures. But, GRU has fewer parameters by dropping the traditional output gate.

b) CRNN-TF: The convolutional recurrent neural network in Time and Frequency dimensions (CRNN-TF) is a variant of CRNN. It can extract spatial dependencies in both the Time and Frequency dimensions of music signals. CRNN-TF has achieved promising performances in several state-of-the-art deep learning-based music models.

E. Multi-instance Attention Mechanism

Given the musical segment embeddings $s_{p,a} = \{s_1, s_2, ..., s_m\}$, we apply a plain selective attention over them to get the musical genre representation q_g for classifying the genres. We adopt q_g as attention query vector initialed by

xavier uniform. The attention for each musical segment in s_k is defined as follows:

$$e_k = \tanh(W_s[s_k; q_g]) + b_s \tag{1}$$

$$a_k = \frac{\exp(e_k)}{\sum_{j=1}^{m} \exp(e_j)} \tag{2}$$

where $[x_1; x_2]$ denotes the vertical concatenation of x_1 and x_2, W_s is the weight matrix, and b_s is the bias. The converged nodes will share the parameters. By doing so, we can compute attention operations on each label of the music segments to obtain their corresponding musical genre representations.

$$g_{p,a} = \mathrm{ATT}(q_g, s_1, s_2, ..., s_m) \tag{3}$$

The musical genre representation g_s will be fed to compute the conditional probability $P(g|h, t, s)$, which is shown below:

$$P(g|h, t, S_{p,a}) = \frac{\exp(o_g)}{\sum_{\hat{g} \in \mathcal{G}} \exp(o_{\hat{g}})} \tag{4}$$

$$o = M g_{p,a} \tag{5}$$

where o is the scores of the music genres. We then use a discriminative matrix M to obtain these genre scores, as Equation 5 shown.

IV. Experiments

In this section, we first introduce the experimental settings. Second, we evaluate our proposed methods and various baseline models in terms of overall and long-tail classification performance. To comprehensively understand how MATT affects the performance of models in identifying the music segment without valid album ID or artist ID, we conduct a case study to explain it. In the end, we introduce the threats that may affect the validity.

A. Experimental Setting

1) Datasets: We evaluate our proposed methods on the FMA dataset which is arranged in a taxonomy of 16 genres. The FMA dataset is a representative dataset used to evaluate the performance of the MGC model [11], [13]. Depending on the number of samples in the dataset, the FMA dataset is available for 4 datasets with different scales, such as full, large, medium, and small datasets. We utilize the 16 classes medium dataset, which demonstrates long-tail distribution, for evaluation.

2) Metrics and Evaluation Procedure: For evaluation, we draw precision-recall curves for all models. To further verify the performance of our model for long-tail genres, we report the Top@K Accuracy results. The baseline dataset and codes can be found in Github[2] [3].

3) Comparison Models: For baselines, we evaluate both neural network-based and feature-based models. To verify the performance of the multiple-instance learning mechanism in music genre classification, we report the results of our method with various baseline models including **LR** [20], **SVM** [21], **KNN** [22], **MLP** [23], **CRNN**, **CRNN-TF**, and the same model equipped with various advanced learning strategies

[2] https://github.com/mdeff/fma

[3] https://github.com/FishInMedi/CRNN-TF

4) Hyperparameter Setting and Reproducibility: To make the results reproducible, we adopt the default data split schema proposed by FMA and the default hyperparameter settings from baseline methods.

B. Overall Evaluation Results

To evaluate the performance of our proposed model, we compare the precision-recall curves and accuracy of our model with various baseline models. The accuracy results are demonstrated in Table 1 and Figure 4s. As shown in both Table 1 and Figure 4(a) and 4(b), we observe that: (1) For the feature engineering-based models, the accuracy of MLP models with MIL and MATT mechanisms outperforms the other baseline models. We also observe similar trends in CRNN and CRNN-TF. These results confirm that the MIL and MATT can significantly improve the model performance. (2) The accuracy of models with MATT is slightly lower than those of models with MIL mechanisms. Notice that the distribution of the dataset is long-tail, so only the accuracy metric cannot help conclude the performance of the models.

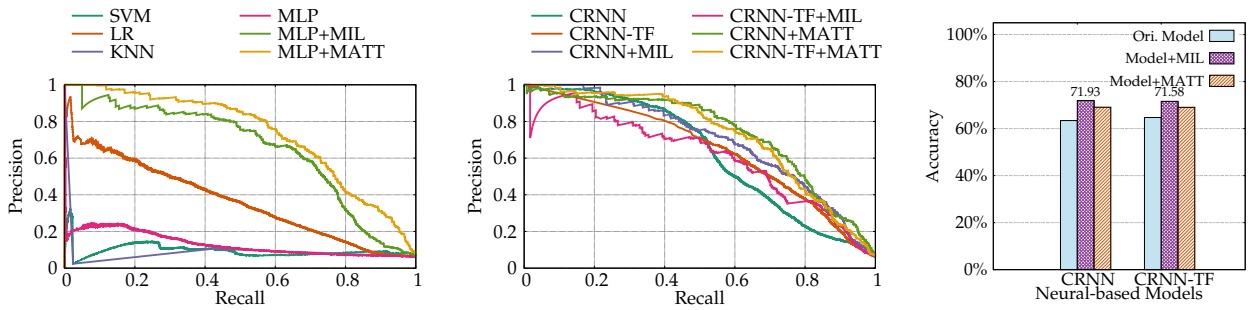
To effectively evaluate the long-tail genres, we adopt the precision-recall curve, which is one of the most widely-adopted evaluation metrics in information retrieval field, to help evaluate the models better.The precision-recall curves are shown in Figure 4(a) and 4(b). We can conclude from both figures:

TABLE I
Testing Accuracy (%) of various features and models on the FMA Data Medium Subset

Feature Set	Dim.	LR	KNN	SVM	MLP	MLP+MIL	MLP+MATT
1 Chroma	84	44	44	48	49	**47.74**	39.53
2 Tonnetz	42	40	37	42	41	**45.90**	43.45
3 MFCC	140	58	55	61	53	**64.71**	63.19
4 Spec. Centroid	7	42	45	46	48	**49.36**	44.93
5 Spec. BW..	7	41	45	44	45	43.26	**44.23**
6 Spec. Contrast	49	51	50	54	53	**55.69**	49.59
7 Spec. Rolloff	7	42	46	48	48	49.16	**49.20**
8 RMS Energy	7	37	39	39	39	**41.08**	38.67
9 Zero-Crossing	7	42	45	45	46	**47.96**	46.91
3+6	189	60	55	63	54	65.68	**68.91**
3+6+4	196	60	55	63	53	**65.84**	63.31
1 to 9	518	61	52	63	58	**69.22**	68.40

(1) All models equipped with MATT mechanism can achieve good precision while their recalls are smaller than 0.6. The feature engineering-based models without equipping any advanced mechanisms show the worst performance among all the models. Besides, the models equipped with MATT mechanism outperform those equipped with MIL. The MATT-equipped MLP and CRNN-TF achieve the best performance under different feature extraction methods. In addition, CRNN-TF achieves the best performance among all the models. (2) The models equipped with MIL mechanism are better than their corresponding baseline models, except for the CRNN-TF. Such phenomena reveal that although the CRNN-TF with MIL achieves state-of-the-art accuracy, the performance of CRNN-TF in music information retrieval benchmarks with long-tail distribution should still be questioned. However, the attention-based CRNN-TF with MIL helps address this issue.

From these experimental results, we can conclude that the MATT-based models achieve the best performances com-



(a) Precision-recall curves of the proposed methods vs feature-based baselines.

(b) Precision-recall curves of the proposed methods vs neural-based baselines.

(c) Testing performance of neural-based models on the FMA Dataset.

Fig. 4. PR-Curve and Accuracy of Our Proposed Models against Baselines

pared with other models. Considering the accuracy, precision, and recall at the same time, CRNN-TF with MATT achieves the best performance.

C. Evaluation Results for Long-tail Genres

TABLE II
TOP@K ACCURACY (%) ON LONG-TAIL CLASSES.

Number of Training Instances		**<100**			**<200**	
Top@K	**K=2**	**K=3**	**K=5**	**K=2**	**K=3**	**K=5**
1 LR	<5.0	7.06	25.29	<5.0	6.38	22.87
2 KNN	17.06	25.88	45.29	15.43	23.94	41.49
3 SVM	11.76	19.41	44.71	10.64	17.55	40.42
4 MLP	<5.0	10.59	28.23	<5.0	10.11	27.13
5 MLP+MATT	**28.24**	**44.71**	**47.06**	**26.60**	**41.49**	**44.15**
6 CRNN	9.41	13.53	34.71	8.51	12.23	38.83
7 CRNN + MATT	**54.12**	**55.88**	**57.65**	**49.47**	**51.60**	**61.7**
8 CRNN-TF	<5.0	9.41	20.59	<5.0	8.51	18.62
9 CRNN-TF + MATT	**54.12**	**63.53**	**65.29**	**59.04**	**61.17**	**71.21**

To further demonstrate the improvements in performance for long-tail genres, we extract a subset of the testing dataset in which all the genres have fewer than 100/200 training instances. We employ the Top@K metric for evaluating the results. For each album-artist pair, the evaluation requires its corresponding golden genre in the first K candidate genres recommended by the models. Since it is difficult for any of the existing models to extract long-tail genres, we select K from 2, 3, 5. We report the micro average Top@K accuracy for these subsets in Table 2. Similar to former contexts, the 1-5 models use the 1-9 feature set in Table 2 as the input, while the 6-9 models adopt log-amplitude mel-spectrogram features as the input. From this table, we have the following observations: (1) Tough the models without MATT achieves high accuracy, their performance in predicting the genres of long-tail music degrades dramatically. It reveals that conventional works, both feature engineering-based and neural-based works, are suffering from an extreme long-tail problem. (2) For both feature engineering-based Models and neural-based models, the models with our MATT outperform other baseline models in extracting long-tail music genres. Among the MATT based works, the CRNN-TF with MATT achieves the best performance. (3) The neural-based models outperform the feature engineering-based models in identifying data-poor genres. It confirms that the neural-based models, which automatically extract music features from the Log-amplitude Mel-spectrogram data, have more potential in long-tail music genre identification than the complex feature engineering-based models. (4) Although our MATT mechanism has achieved better results in the task of long-tail music genre classification as compared with other SOTA methods, the results of all the ML and DL algorithms with MATT are still not satisfying. In the near future, we plan to adopt more advanced schemes [24]–[26] and introduce extra information to solve this problem.

D. Case Study

We noticed that the bag-level evaluation requires extra information, such as album ID and artist ID. However, in some scenarios, this information is hard to obtain. Thus we carry out a case study on segment-level evaluation. The segment-level evaluation does not need any extra information but the music segment in the evaluation process. In the training phase, since the album ID and artist ID is easy to obtain, we train the models using the processed bag-level dataset. In the testing phase, we evaluate the models without using the multi-instance learning strategy. In other words, the model is only equipped with a plain attention mechanism in the evaluation stage.

We conduct experiments using MLP with the whole FMA feature sets. By training the MLP model on the bag-level dataset using features 1 to 9 from the feature set, the testing accuracy degrades from 68.83% to 63.69%. Nevertheless, we argue that this degradation in testing accuracy is acceptable. Because in the testing phase, no album ID or artist ID is used to help identify the genre information. In addition, we found that the testing accuracy is still higher than that of the training model at the segment level, whose testing accuracy is only 58%. For the rest features from the feature set, we observed a similar trend in terms of accuracy and average precision.

E. Threats to Validity

1) ML Platform: We implement MATT on an industrial-grade ML platform, Pytorch [27]. We notice the baseline works were implemented in Scikit-learn [28] and Keras [29]. For the metrics not reported in the baseline works, we extract the parameters of the models and try our best to keep the parameters the same as before. For the metrics has been reported in previous works, we adopt the results from baseline works to avoid the possible evaluation bias.

2) Hardware: We notice that different hardware may cause differences in the performance of music genre classification models, so we declare the hardware we use in our experiments. In our study, we use a Linux server with two 48-core Intel CPUs and 376 GB of memory. The machine is also equipped with four NVIDIA RTX 3090 GPUs.

V. Conclusion

In this paper, we propose the MATT mechanism to identify music genres, especially the long-tail genres, in a large-scale music benchmark. Comprehensive experimental results demonstrate that the MATT can significantly improve the performance of the MGC models for identifying long-tail music genres. Moreover, MATT can enhance the models' capability of classifying music genres in both feature engineering- and neural network-based music genre classification models. Evaluation results on the segment-level testing dataset also demonstrate that the MATT is still competitive even when album ID and Artist ID are missed.

References

[1] D. C. Corrêa and F. A. Rodrigues, "A survey on symbolic data-based music genre classification," *Expert Syst. Appl.*, vol. 60, pp. 190–210, 2016.

[2] C. Sénac, T. Pellegrini, F. Mouret, and J. Pinquier, "Music feature maps with convolutional neural networks for music genre classification," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI 2017, Florence, Italy, June 19-21, 2017*. ACM, 2017, pp. 19:1–19:5.

[3] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in neural information processing systems*, vol. 22, 2009.

[4] Z.-H. Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, vol. 1, 2004.

[5] X. Liu, F. Zhao, X. Gui, and H. Jin, "Lekan: Extracting long-tail relations via layer-enhanced knowledge-aggregation networks," in *International Conference on Database Systems for Advanced Applications*. Springer, 2022, pp. 122–136.

[6] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and adaboost for music classification," *Machine learning*, vol. 65, no. 2, pp. 473–484, 2006.

[7] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 282–289.

[8] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," vol. 1, pp. 113–116, 2002.

[9] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 2014, pp. 6959–6963.

[10] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *2015 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 121–125.

[11] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[12] S. S. Ghosal and I. Sarkar, "Novel approach to music genre classification using clustering augmented learning method (calm)." in *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, 2020.

[13] Z. Wang, S. Muknahallipatna, M. Fan, A. Okray, and C. Lan, "Music classification using an improved CRNN with multi-directional spatial dependencies in both time and frequency dimensions," in *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 2019, pp. 1–8.

[14] R. Yang, L. Feng, H. Wang, J. Yao, and S. Luo, "Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices," *IEEE Access*, vol. 8, pp. 19 629–19 637, 2020.

[15] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2021.

[16] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning for audio-based music classification and tagging," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 67–74.

[17] S. Craw, B. Horsburgh, and S. Massie, "Music recommendation: Audio neighbourhoods to discover music in the long tail," in *International Conference on Case-Based Reasoning*. Springer, 2015, pp. 73–87.

[18] V. D. Valerio, R. M. Pereira, Y. M. G. Costa, D. Bertolini, and C. N. S. Jr., "A resampling approach for imbalanceness on music genre classification using spectrograms," in *Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference*, K. Brawner and V. Rus, Eds. AAAI Press, 2018, pp. 500–505.

[19] X. Han, P. Yu, Z. Liu, M. Sun, and P. Li, "Hierarchical relation extraction with coarse-to-fine grained attention," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2236–2245.

[20] R. E. Wright, "Logistic regression." 1995.

[21] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification*. Springer, 2016, pp. 207–235.

[22] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[25] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[26] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] F. Chollet *et al.* (2015) Keras.