Quantifying uncertainty in music genre classification

Hanna Lukashevich[1], Sascha Grollmisch[1], Jakob Abeßer[1]
[1] *Fraunhofer IDMT, 98693 Ilmenau, Deutschland, Email: hanna.lukashevich@idmt.fraunhofer.de*

Abstract
Music annotation algorithms apply signal processing and machine learning techniques to extract and add metadata to audio recordings in music archives. One common task is music genre classification, where a single label (such as Rock, Pop, or Jazz) is assigned to each song. Since music genres are often ambiguous, classification algorithms naturally cannot obtain fully correct predictions. Therefore, our focus is not only on the label with the highest posterior class probability but also on on a realistic confidence value for each of the possible genres. In theory, most state-of-the-art classification algorithms based on deep neural networks suffer from overconfident predictions which complicates the interpretation of the final output values. With this work, we investigate whether the problem of overconfident predictions, and therefore non-representative confidence values, is also applying to music genre classification. Furthermore, we outline state-of-the-art methods for preventing this behaviour and investigate the influence of the so-called temperature scaling to get more realistic confidence outputs which can be directly used in real-world music tagging applications.
This study was supported by the German Research Foundation (AB 675/2-2).

Introduction
Over the past few decades, the rapid digitization of music has led to a massive increase in the number of audio recordings available, which made it necessary to develop efficient and effective tools for annotating and organizing these vast archives. In this context, music annotation algorithms have emerged as a prominent area of research since they combine signal processing and machine learning techniques to extract and add metadata to audio recordings, and thereby improve and facilitate the search, discovery, and analysis of music. One pivotal task in this domain is music genre classification, wherein algorithms attempt to assign a single, representative label (e.g., Rock, Pop, or Jazz) to each song based on its characteristics.

Since music genres are inherently ambiguous with many songs exhibiting features of multiple genres, providing accurate predictions is challenging for classification algorithms. Consequently, it is important to not only identify the most probable genre label but also ascertain a realistic confidence value for each potential genre. State-of-the-art classification algorithms, particularly those based on deep neural networks, have been found to suffer from overconfident predictions, which complicates the interpretation of the final output values and hinders the overall efficacy of the classification process [1, 2, 3].

In this work, we explore overconfident predictions and their impact on the reliability of confidence values in music genre classification. We delve into state-of-the-art methods designed to mitigate overconfidence in predictions and examine the effectiveness of temperature scaling. Temperature scaling calibrates classifier outputs to generate more realistic confidence values. We aim to contribute to the development of music tagging applications by investigating the influence of temperature scaling on the performance of music genre classification algorithms. This way the algorithms become more robust and reliable and thus can be deployed in real-world scenarios.

Related Work
The output of the last softmax layer is often erroneously interpreted as the confidence of a model with respect to the class decision. This interpretation problem is called *deterministic overconfidence* and is caused by obtaining point estimates of activations instead of distributions of estimates [1]. As a result, not only the correct but also the erroneous class decisions are getting high softmax outputs. Hein et al. [3] show that the deterministic overconfidence is large when the data is far away from the model's decision boundary or when rectified linear units (ReLU) are used in the network.

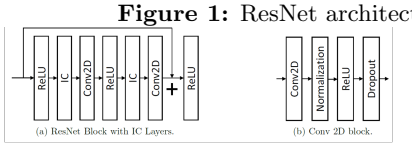
Numerous strategies have been developed to address the issue of deterministic overconfidence in classification algorithms. One such approach, known as *temperature scaling*, involves calibrating the softmax outputs post-hoc to soften prediction outcomes [2]. This technique is accomplished by dividing the output logits of the neural network by a fixed temperature value T. Temperature scaling effectively mitigates deterministic overconfidence, particularly when data can be regarded as in-distribution. Alternative approache to address overconfidence is Monte Carlo (MC)-Dropout technique [1], which introduces dropout layers to neural networks to model uncertainty and efficiently approximate Bayesian inference in deep Gaussian processes. During the inference process, the dropout layers remain active, and the same input passes through the neural network multiple times. Uncertainty can also be estimated using deep ensembles, which consist of multiple independently trained neural networks [4]. This method leverages the collective knowledge and diversity of the ensemble members to provide a more comprehensive understanding of the data, thereby reducing the impact of overconfidence in the model's predictions.

Dataset
The FMA (Free Music Archive) small dataset is a curated subset of the larger FMA dataset, specifically designed

for research in music genre classification [5]. It contains 8,000 tracks spanning 8 genres, including Classical, Hip-Hop, Electronic, Folk, Rock, Experimental, International, and Instrumental. Each track is 30 seconds long and comes with metadata such as artist, album, and track information. It contains pre-defined balanced training, validation, and evaluation splits.

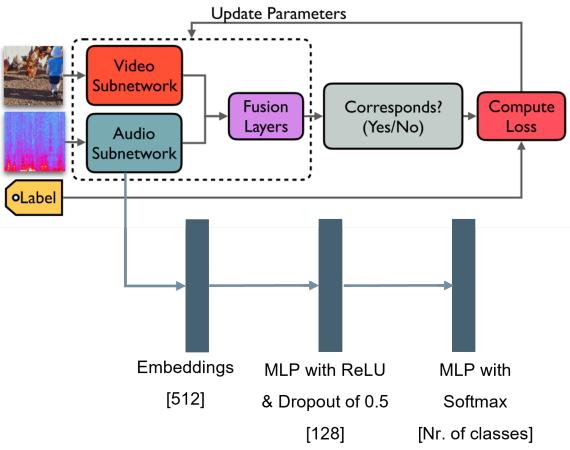
This paper does not include a comprehensive overview of the state of the art on the FMA small dataset. For the sake of comparison, we only discuss here a few examples of the recent publications for this dataset. Zhao et al. [6] report an accuracy of 56.4% using a self-supervised pre-training method with a Swin Transformer, which leverages massive unlabeled music data to enhance music classification performance and diminish reliance on extensive labeled music datasets. Kostrzewa et al. [7] compare multiple deep learning network architectures, including Convolutional Neural Network (CNN), 1-Dimensional Convolutional Recurrent Neural Network (CRNN), 2-Dimensional CRNN, Recurrent Neural Network with Long Short-Term Memory (LSTM) Cells and ensembles of stacked CNNs and CRNN variants. The highest single model accuracy of 51.63% was achieves with a CNN and the ensembles were able to improve the accuracy to 56.39%. Kostrzewa et al. [7] also bring an overview of the latest publications on FMA small dataset with the corresponding accuracies.

Figure 1: ResNet architecture



Layer	Output	Kernel Size	Dropout
Conv 2D	64	(5, 5)	-
Relu	-	-	-
ResNet Block	64	(3, 1)	0.10
Avg. Pooling	-	(2, 2)	-
ResNet Block	64	(3, 3)	0.10
Avg. Pooling	-	(2, 2)	-
ResNet Block	64	(3, 3)	0.10
ResNet Block	128	(3, 1)	0.10
Avg. Pooling	-	(2, 2)	-
ResNet Block	256	(1, 1)	0.10
Avg. Global Pooling	-	-	-
Softmax	Nr. of classes	-	-

Figure 2: OpenL3+MLP architecture



Experiments

In the present study, our primary objective is not to identify the optimal deep learning architecture for automatic music genre classification. Instead, we focus on examining the posterior class probabilities and exploring the impact of temperature scaling to obtain more realistic confidence outputs. These calibrated outputs have the potential to be directly employed in real-world music tagging applications.

To conduct our experiments, we have selected two distinct network architectures. The first architecture is a Residual Network (ResNet) consisting of 420 thousand parameters, as illustrated in Figure 1 and described in detail in Grollmisch et al. [8]. The second architecture is a shallow Multi-Layer Perceptron (MLP) built on top of the well-established OpenL3 embeddings [9], shown in Figure 2. Throughout the remainder of this paper, we will refer to the first architecture as ResNet and the second architecture as OpenL3+MLP.

The ResNet architecture processes 3-second patches of mel spectrogram with 96 mel bands as input. These patches are extracted with a hop size of 1 second. The mel spectrogram is generated from the 44.1kHz audio signal using a window size of 2048 and a hop size of 710. The OpenL3+MLP architecture employs a standard OpenL3 audio input and generates an output consisting of 512 embedding dimensions.

Our results revealed a mean filewise accuracy of 49.12% for the ResNet architecture and a mean filewise accuracy of 53.6% for the OpenL3+MLP architecture. The confusion matrix for the ResNet model is presented in Figure 3. It is evident that the accuracies for distinct classes vary significantly. For instance, the Hip-Hop and Electronic genres exhibit relatively high accuracies of 85% and 64%, respectively, while the Pop and Folk genres demonstrate considerably lower accuracies of 31% and 14%, respectively. The highest confusion is observed for the Folk genre, with 49% being predicted as Experimental. It is worth noting that the random baseline for eight classes is 12.5%.

Similarly, the confusion matrix for the OpenL3+MLP model, as shown in Figure 4, displays varying accuracies across distinct classes. The OpenL3+MLP model achieves higher accuracies for the Hip-Hop and Electronic genres, at 92% and 70%, respectively. However, the accuracies for the Pop and Folk genres remain relatively low, at 29% and 17%, respectively. The similarities observed in the confusion matrices for both models can be attributed to the inherent difficulties that humans also face when attempting to distinctly define certain musical genres. This challenge consequently leads to noisy human annotations present within the dataset.

Further we investigated whether the softmax layer output of our models reflected the true posterior class probabilities for our dataset. Based on existing research and theoretical understanding, we expected discrepancies between the softmax layer output, often referred to as "confidence," and the expected accuracy for a

Figure 3: Confusion matrix for ResNet

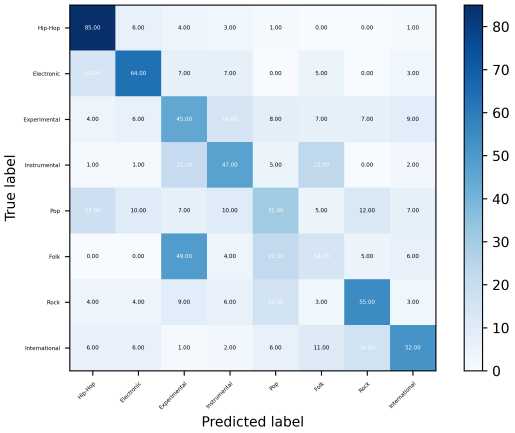


Figure 4: Confusion matrix for OpenL3+MLP

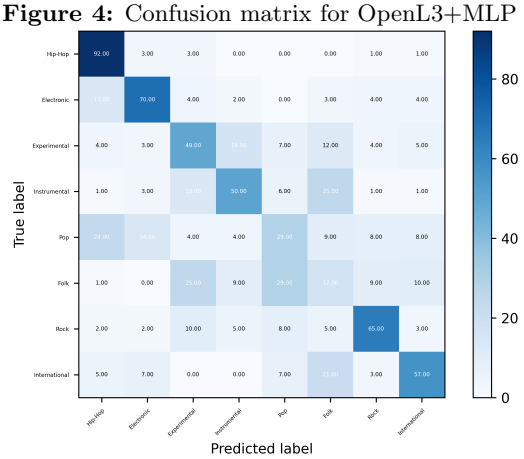


Figure 5: Reliability diagram for ResNet

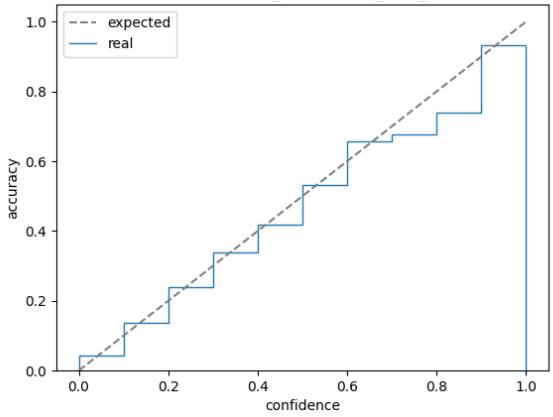
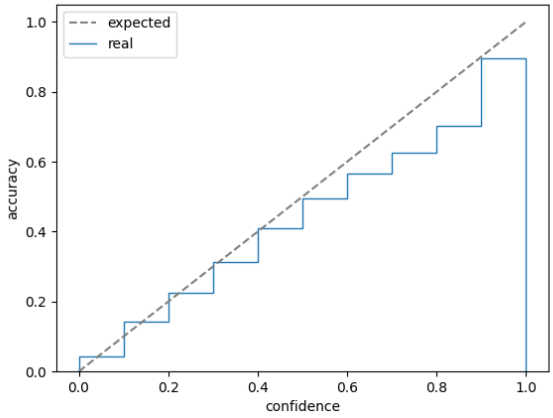


Figure 6: Reliability diagram for OpenL3+MLP



specific classifier decision. To analyze these differences, we constructed ten data buckets for ten intervals of confidence levels. This procedure was repeated for both ResNet and OpenL3+MLP models. The first data bucket contained all FMA small test items with confidence values between 0.9 and 1.0. Given the high confidence values, we expected the classification accuracy for these items to be high, averaging around 95%. The next bucket included test items with confidences between 0.8 and 0.9, where we expected an average accuracy of 85%, etc. We continued this process of forming data buckets with a confidence step of 0.1 until the last bucket, which contained items with confidences between 0.0 and 0.1. For this final bucket, only 5% of the class decisions were expected to be correct, corresponding to an expected average accuracy of 5%.

Our analysis led to the creation of reliability diagrams, shown in Figure 5 for the ResNet model and in Figure 6 for the OpenL3+MLP model. The dashed grey line in these figures represents the expected accuracy, while the blue step function indicates the actual accuracy for each bucket.

The results revealed that both models are poorly calibrated and exhibit the expected deterministic overconfidence. For example, the data items with confidences between 0.8 and 0.9, which has a mean expected accuracy of 85%, actually exhibit lower mean accuracies of approximately 74% for the ResNet model and 70% for the OpenL3+MLP model. Given that the deterministic overconfidence is more pronounced for the OpenL3+MLP model, and this model also has a higher overall accuracy, we will focus our discussion on the results for this particular model.

As the next step, we experimented with temperature scaling and found that it provides a powerful tool for calibrating the softmax layer output. Figure 7 displays the reliability diagram for the OpenL3+MLP model with temperature scaling using $T = 1.25$, $T = 1.50$, and $T = 1.75$. The optimal value for temperature scaling can be chosen using the validation part of the dataset.

Even if the softmax layer output is well calibrated with a suitable temperature scaling, there is still one major issue remaining. As we have seen from the confusion matrices in Figure 3 and Figure 4, the accuracies for musical genres differ strongly. We investigated the differences for the musical genres in the reliability diagrams. Considering that, we further split the buckets for the reliability diagram according to the musical genres. The results for the OpenL3+MLP model are shown in Figure 8. The diagram shows that the classifer is extremely overconfident for the poorly performing classes Folk and Pop. In contrast, the class decisions for some of the better performing classes tend to output confidences

Figure 7: Reliability diagram for OpenL3+MLP with temperature scaling

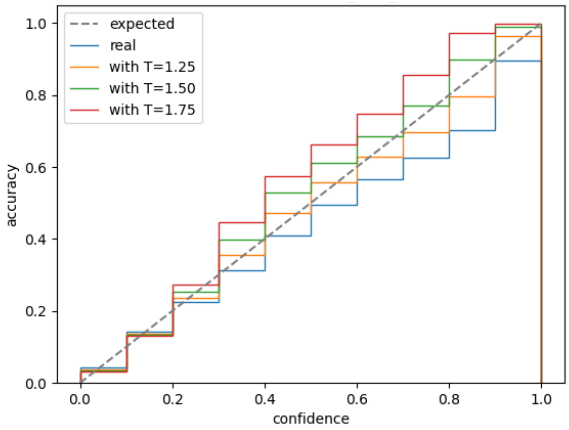
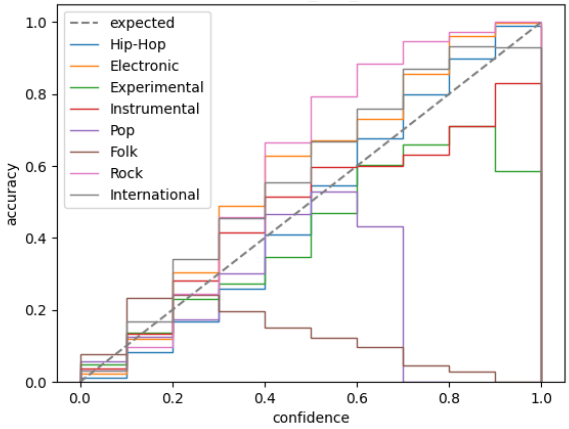


Figure 8: Reliability diagram per class for OpenL3+MLP with temperature scaling at $T = 1.25$



higher that are lower than real accuracy values for the bucket.

Conclusion

This study has highlighted several critical aspects related to automatic music genre classification. While current methods have made significant advancements in this field, automatic music genre classification remains an imperfect task. Our analysis has shown that state-of-the-art deep learning approaches, despite their remarkable performance, are not entirely reliable when it comes to estimating posterior class probabilities for music genre classification. To address this issue, we have explored the use of temperature scaling, which calibrates the softmax output and can improve the reliability of probability estimates. However, it is essential to recognize that the optimal value for temperature scaling is both dataset and model dependent, requiring careful selection and tuning.

Furthermore, we have observed that posterior class probabilities for different music genres are heterogeneous and require extra attention and individual calibration. This finding underscores the need for more robust uncertainty estimation techniques tailored to the specific characteristics of each music genre.

To enhance the uncertainty estimation process, we recommend investigating additional methods such as Monte-Carlo Dropout and deep ensembles. These techniques have the potential to further improve the reliability and interpretability of automatic music genre classification systems.

In summary, our work contributes to the ongoing development of automatic music genre classification by emphasizing the importance of uncertainty estimation and exploring potential solutions. The incorporation of these insights into future research will undoubtedly lead to more accurate, reliable, and useful music genre classification systems.

References

[1] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. of International conference on machine learning (ICML)*, 2016, pp. 1050–1059.

[2] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. of International conference on machine learning (ICML)*, 2017, pp. 1321–1330.

[3] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[5] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proc. of the 18th International Society for Music Information Retrieval Conference*, 2017.

[6] H. Zhao, C. Zhang, B. Zhu, Z. Ma, and K. Zhang, "S3T: Self-supervised pre-training with swin transformer for music classification," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 606–610.

[7] D. Kostrzewa, P. Kaminski, and R. Brzeski, "Music genre classification: Looking for the perfect network," in *Proc. of the 21st International Conference in Computational Science (ICCS)*, 2021, pp. 55–67.

[8] S. Grollmisch and E. Cano, "Improving semi-supervised learning for audio classification with fixmatch," *Electronics*, vol. 10, no. 15, p. 1807, 2021.

[9] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.