How reliable are posterior class probabilities in automatic music classification?

Hanna Lukashevich
Fraunhofer Institute for Digital Media Technology IDMT
Ilmenau, Germany
hanna.lukashevich@idmt.fraunhofer.de

Sascha Grollmisch
Fraunhofer Institute for Digital Media Technology IDMT
Ilmenau, Germany
sascha.grollmisch@idmt.fraunhofer.de

Jakob Abeßer
Fraunhofer Institute for Digital Media Technology IDMT
Ilmenau, Germany
jakob.abesser@idmt.fraunhofer.de

Sebastian Stober
Otto-von-Guericke-University Magdeburg
Magdeburg, Germany
stober@ovgu.de

Joachim Bös
Fraunhofer Institute for Digital Media Technology IDMT, Technische Universität Ilmenau
Ilmenau, Germany
joachim.boes@idmt.fraunhofer.de

ABSTRACT

Music classification algorithms use signal processing and machine learning approaches to extract and enrich metadata for audio recordings in music archives. Common tasks include music genre classification, where each song is assigned a single label (such as Rock, Pop, or Jazz), and musical instrument classification. Since music metadata can be ambiguous, classification algorithms cannot always achieve fully accurate predictions. Therefore, our focus extends beyond the correctly estimated class labels to include realistic confidence values for each potential genre or instrument label. In practice, many state-of-the-art classification algorithms based on deep neural networks exhibit overconfident predictions, complicating the interpretation of the final output values. In this work, we examine whether the issue of overconfident predictions and, consequently, non-representative confidence values is also relevant to music genre classification and musical instrument classification. Moreover, we describe techniques to mitigate this behavior and assess the impact of deep ensembles and temperature scaling in generating more realistic confidence outputs, which can be directly employed in real-world music tagging applications.

CCS CONCEPTS

• **General and reference** → *Reliability*; • **Information systems** → *Relevance assessment*; **Content analysis and feature selection**.

KEYWORDS

music information retrieval, music classification, uncertainty, temperature scaling, deep ensembles

ACM Reference Format:
Hanna Lukashevich, Sascha Grollmisch, Jakob Abeßer, Sebastian Stober, and Joachim Bös. 2023. How reliable are posterior class probabilities in automatic music classification?. In *Audio Mostly 2023 (AM '23), August 30–September 01, 2023, Edinburgh, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3616195.3616228

AM '23, August 30–September 01, 2023, Edinburgh, United Kingdom

ACM ISBN 979-8-4007-0818-3/23/08.
https://doi.org/10.1145/3616195.3616228

INTRODUCTION

In recent decades, the volume of available audio recordings has grown tremendously due to the swift digitization of music. As a consequence, efficient and effective tools are needed to annotate and organize music archives. Automatic music annotation has emerged as a significant research area, where signal processing and machine learning methods are combined to extract metadata from audio recordings. Thus, music annotation allows for enhancing and simplifying the search, discovery, and analysis of music.

A crucial task within Music Information Retrieval (MIR) is music genre classification, in which algorithms attempt to assign a single, representative label (e.g., Rock, Pop, or Jazz) to summarize the musical style of a song. Music genres are inherently ambiguous since songs from multiple genres often share similar features. As a consequence, delivering accurate predictions poses a challenge for classification algorithms. Therefore, it is essential to not only determine the most likely genre label, but also establish a realistic confidence value that goes along with it.

Another prevalent MIR task is instrument classification, which involves detecting a single instrument or an instrument family in monophonic or polyphonic recordings. The main challenges stem from the timbral similarities between instruments within the same instrument families, such as viola and violin, and various models that exist for the same type of instrument, like electric and acoustic guitars. In addition, music recordings from different eras were produced using various recording and mixing techniques. Therefore, a realistic confidence output for each detected instrument is of interest.

Deep learning-based classifiers are the state of the art for both retrieval tasks [6, 8, 10, 16, 17]. However, these classifiers have been observed to generate overconfident predictions, which complicates the interpretation of the final output values and undermines the overall effectiveness of the classification process [5, 7, 9].

As the main contribution of this study, we investigate the reliability of confidence values in the classification of musical genres

and instrument families using neural network-based classifiers. We propose the mean absolute error to quantify the overconfidence or underconfidence of their outputs. We evaluate whether the state-of-the-art methods temperature scaling and deep ensembles allow for obtaining more realistic posterior class probabilities. Temperature scaling rescales classifier outputs, while deep ensembles comprise multiple independently trained neural networks that collectively generate the final output. Our goal is to improve music classification systems by building more reliable classifiers for real-world applications. This work builds on our preliminary work [12] by extending the results from music genre to instrument family classification, including deep ensembles in addition to temperature scaling, and adding several important contributions in this field.

RELATED WORK

The uncertainty of the classification decision can be quantified using a confidence measure which is a score that accompanies the decision and signifies its trustworthiness. A higher confidence corresponds to a more reliable decision.

The importance of confidence arises when it is necessary **(i)** to compare or merge classification decisions from different classifiers, **(ii)** to implement a reject option based on the confidence, or **(iii)** to interpret classification outcomes.

In this paper, we define confidence as a value ranging from 0 to 1 that is associated with a classification decision and meets the criteria set forth by Duin and Tax [3]:

(1) On average, a proportion c of all objects with a confidence of c should be classified accurately.
(2) Objects that are classified reliably should possess higher confidences than objects near the decision boundary.

Confidences of this nature are simple to understand. For example, if we obtain 100 decisions with confidences around 0.7, we can anticipate approximately 70 of them to be accurate.

For multi-class single-label tasks, the output of the last layer with a softmax activation function is often erroneously interpreted as the confidence of a model with respect to the class decision. The divergence from reliable posterior class probabilities is caused by obtaining point estimates of activations instead of distributions of estimates, a phenomenon referred to as *deterministic overconfidence* [5]. As a result, the output probability of the winning class is often higher than it should be. Furthermore, not only the correct but also the erroneous class decisions are getting high softmax outputs, which complicates the implementation of the reject option. Hein et al. [9] 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.

Various strategies have been developed to address the issue of deterministic overconfidence in classification algorithms: *Temperature scaling* involves calibrating the softmax outputs post-hoc to soften or sharpen the output posterior class probabilities [7]. This technique is accomplished by dividing the output logits of the neural network by a fixed temperature value T before being passed to the softmax function. Temperature scaling with values larger than 1 effectively mitigate deterministic overconfidence, particularly when data can be regarded as in-distribution. In-distribution and out-of-distribution refer to the relationship between training and testing data in machine learning. In-distribution data is representative of the training set and follows the same underlying patterns or features. Out-of-distribution data deviates from the training set, containing novel or unexpected instances that the model may not have encountered during training, thus challenging its generalization capabilities.

As an alternative approach to reduce deterministic overconfidence, Monte Carlo (MC)-Dropout [5] introduces dropout layers to model uncertainty and to 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, leading to slightly different results on each pass. As a third approach, uncertainty can also be estimated using deep ensembles, which consist of multiple independently trained neural networks [11]. 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.

Deep ensembles have been demonstrated to surpass MC-Dropout in terms of quantifying uncertainty across various datasets and tasks in both regression and classification [11]. Moreover, deep ensembles have established themselves as state-of-the-art in out-of-distribution settings, such as data perturbations or the introduction of novel classes not seen during training [13]. The advantage of deep ensembles in out-of-distribution settings can be attributed to the significant differences in their weight values and loss trajectories, which result in diverse predictions. Investigations in [11] and [13] indicate that just a few independently trained neural networks from the ensemble—specifically, just five models—are sufficient for achieving these results.

DATASET

In this study we address music genre classification and instrument family classification with the Free Music Archive (FMA) [2] and NSynth datasets [4], respectively. For the sake of comparison, we only discuss here a few examples of the recent publications using these datasets.

The FMA small dataset is a carefully selected portion of the more extensive FMA dataset, created specifically for studying music genre classification. This dataset features 8 000 tracks across eight genres: Classical, Hip-Hop, Electronic, Folk, Rock, Experimental, International, and Instrumental. Each track has a duration of 30 seconds and is accompanied by metadata such as artist, album, and track details. The dataset includes predefined balanced splits for training, validation, and evaluation. Zhao et al. [17] achieved a 56.4% accuracy using a self-supervised pre-training approach with a Swin Transformer, which takes advantage of large volumes of unlabeled music data to improve music classification results and reduce dependence on sizable labeled music datasets. Kostrzewa et al. [10] compared various deep learning network architectures, such as Convolutional Neural Networks (CNNs), 1-Dimensional Convolutional Recurrent Neural Networks (CRNNs), 2-Dimensional CRNNs, Recurrent Neural Networks with Long Short-Term Memory (LSTM) cells, and ensembles of stacked CNNs and CRNN variants. The highest single-model accuracy of 51.63% was achieved with

a CNN, and ensembles of several CNNs increased the accuracy to 56.39%.

The NSynth dataset contains 300k musical notes sampled from over 1k instruments. These instruments belong to 10 instrument families such as Bass, Brass, and String[1]. It comes with a separate test set that contains only unseen instruments for each family. The files were recorded with a sample rate of 16 kHz and a duration of four seconds. Current supervised state-of-the-art methods reported a classification accuracy of 74.7% using a CNN in combination with audio effects such as chorus and flanger [14]. Saeed et al. [15] reported 73.0% with pre-trained Contrastive Learning for Audio (COLA) embeddings. Grollmisch and Cano [6], obtained 77.1% with a Residual Network (ResNet)-based CNN and random image augmentations of the log mel spectrograms.

EXPERIMENTAL PROCEDURE

In this work, our primary objective is not to identify the optimal deep learning architecture for automatic music genre and instrument family classification. Instead, we focus on examining the posterior class probabilities and exploring the impact of temperature scaling and deep ensembles to obtain more realistic confidence outputs.

For our experiments, we chose two distinct network architectures. The initial architecture is a ResNet comprising 420k parameters and elaborated in [6]. The subsequent architecture is a shallow Multi-Layer Perceptron (MLP) constructed atop the widely-recognized OpenL3 embeddings [1]. In the rest of this paper, we will refer to the first architecture as ResNet and the latter as OpenL3. As in [6], the ResNet is trained with random image augmentations applied to the mel spectrogram. All classifiers are trained with Adam optimizer and a learning rate of 10^{-3} for 100 epochs.

On FMA, the ResNet processes 3-second patches of log mel spectrogram extracted with 96 Mel bands, a window size of 2048 samples, and a hop size of 710 samples. For NSynth, the 4 second long audio files form one log mel spectrogram input patch, which is extracted with 64 Mel bands, a window size of 2048 samples, and a hop size of 1024 samples. For OpenL3 we use the audio branch trained with music data and an embedding size of 512 values.

In the conducted experiments, the ResNet and OpenL3 models are trained on the corresponding training subsets of the FMA small and NSynth datasets and tested on the respective test subsets. During the inference stage, the outputs of the softmax layers are taken as a proxy of posterior class probabilities for each patch, followed by averaging these estimates over all patches in each file. This methodology relies on the assumption that the musical genre or played musical instrument remain consistent within each file in the datasets, based on the fact that labels are only provided for the whole recording.

As suggested in [11], the networks are trained five times with the random initialization to form a deep ensemble. The posterior class probabilities for the ensemble are calculated as the mean over the output of the single models in the ensemble. For temperature scaling, the logits are multiplied with the temperature value before being passed to the softmax activation, which outputs the adjusted class probabilities.

[1]Since the test dataset contains no recordings for `synth_lead` we exclude this instrument family from our experiments.

	Dataset	Architecture	Accuracies in %
Single models	FMA	ResNet	47.22 (0.78)
Ensemble	FMA	ResNet	50.74
Single models	FMA	OpenL3	45.57 (0.18)
Ensemble	FMA	OpenL3	46.70
Single models	NSynth	ResNet	79.96 (0.61)
Ensemble	NSynth	ResNet	81.49
Single models	NSynth	OpenL3	63.99 (0.17)
Ensemble	NSynth	OpenL3	65.32

Table 1: Accuracy values for both datasets and network architectures in %. The accuracy values for single models are provided as mean over all single models with the standard deviation in parentheses.

.

RESULTS

The classification accuracy values for single models as well as for ensembles are presented in Table 1. Notably, the ensembles consistently enhance the accuracy. For example, in the case of the FMA dataset using the ResNet architecture, the ensemble's accuracy exhibits an improvement from a mean accuracy of 47.22 for single models to 50.74 for the ensemble.

As the focus of this work is not on achieving an optimal classification accuracy, but on quantifying the reliability of the classification, we proceeded to examine whether the softmax layer output of our models accurately represents the true posterior class probabilities for our datasets and models. Prior research and theoretical knowledge suggest a potential discrepancy between the softmax layer output, commonly known as "confidence", and the expected accuracy for a specific classifier decision.

In order to investigate these discrepancies, we apply the following analysis. We set ten data buckets corresponding to ten intervals of confidence levels. The initial data bucket encompasses all items with confidence values within the range of 0.9 to 1.0. Considering the high confidence values, we expect the classification accuracy for these items to be high, averaging approximately 95%. The subsequent data bucket contains test items with confidences between 0.8 and 0.9, for which we projected an average accuracy of 85%. We continue generating data buckets in steps of 0.1 confidence until reaching the final bucket, comprising items with confidences between 0.0 and 0.1. For this last bucket, only 5% of the class decisions are expected to be accurate, equating to an expected average accuracy of 5%. This procedure is repeated for each combination of the dataset and the model, "FMA–ResNet", "FMA–OpenL3", "NSynth–ResNet", "NSynth–OpenL3", leading to the creation of four reliability diagrams, shown in Figure 1. The diagonal dashed green lines in these figures represent the expected accuracy, while the stair plots indicate the actual accuracy for each bucket. The light gray lines show the results for single models. The thick black line corresponds to the ensemble of single models. The meaning of the blue lines will be introduced below.

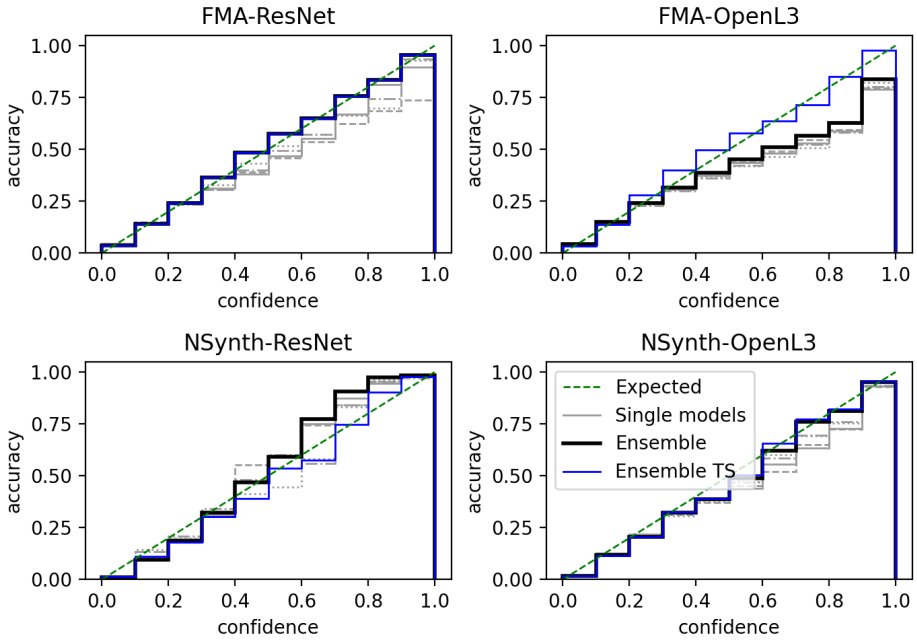

Figure 1: Reliability diagrams for all datasets and models

The results in Figure 1 demonstrate the poor calibration of single models (depicted as grey lines) and exhibit deterministic overconfidence for three of the four experiments ("FMA–ResNet", "FMA–OpenL3", and "NSynth–OpenL3") with "NSynth–ResNet" being an underconfident exception. For instance, in the "FMA–ResNet" experiment, the data items with confidences between 0.8 and 0.9, which should possess a mean expected accuracy of 85%, actually demonstrate lower mean accuracy values of approximately 69% for a single model depicted with a grey dashed line. This means that predictions in this interval are less correct than expected and the model is, therefore, overconfident.

The reliability diagrams for the ensembles are shown in Figure 1 as thick black lines. Ideally, we expect the reliability plots to be as close as possible to the main diagonal, depicted as green dashed lines. Figure 1 shows that the ensemble thick black curves come closer to the main diagonal compared to the grey curves of single models. This effect is especially pronounced for "FMA–ResNet" and "NSynth–OpenL3".

In addition to the visual analysis of reliability diagrams, we use the quantitative analysis and compute the mean absolute error (MAE) between the reliability curves and the expected accuracy values for the confidence buckets. Table 2 shows the MAE values for both datasets and networks. For single models the mean and standard deviation (std) of MAEs are presented. For all experiments except of "NSynth–ResNet", the MAE is considerably lower for the ensembles compared to the single models, confirming the theoretical background outlined in related work. The lowest MAE was obtained for "FMA–ResNet" ensembles, where the reliability is close to the main diagonal, see Figure 1.

However, the results show that the ensembles are not completely solving the issue of unreliable confidence outputs. Therefore, we additionally apply temperature scaling in order to calibrate the confidence values.

Figure 2 shows how the MAE values change depending on the temperature T. For the "FMA–ResNet" configuration, the lowest MAE values for single models are observed at $T = 1.2$, while the ensemble's lowest MAE of 0.013 is achieved at $T = 1.0$. In the case of the "FMA–OpenL3" configuration, there are only minor differences between the MAE curves for single models, which can be attributed to the deterministic behavior of the pre-trained OpenL3 component and the shallow MLP. We observe the minimum ensemble MAE value of 0.026 at $T = 1.65$. The "NSynth–ResNet" experiment is the only instance that exhibits underconfidence. Here we have the

	Dataset	Architecture	MAE
Single models	FMA	ResNet	0.057 (0.015)
Ensemble	FMA	ResNet	0.013
Single models	FMA	OpenL3	0.106 (0.003)
Ensemble	FMA	OpenL3	0.087
Single models	NSynth	ResNet	0.058 (0.006)
Ensemble	NSynth	ResNet	0.068
Single models	NSynth	OpenL3	0.058 (0.007)
Ensemble	NSynth	OpenL3	0.034

Table 2: Mean absolute error (MAE) values for both datasets and network architectures. The MAE values for five single models are provided as the mean over all single models with the standard deviation in parentheses.

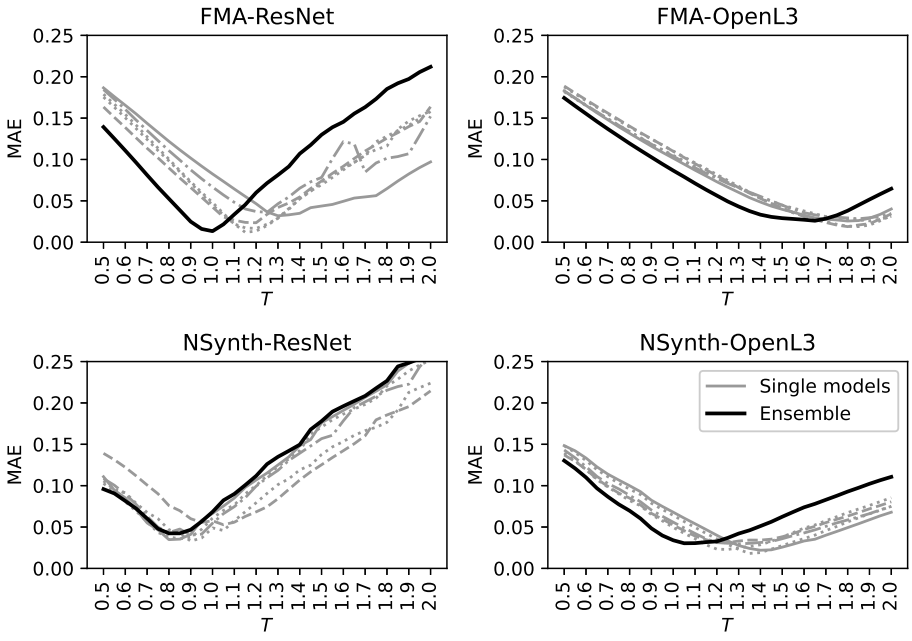

Figure 2: Mean absolute error (MAE) values dependency on temperature scaling (T) for all datasets and models

lowest MAE value of 0.042 at $T = 0.80$. For the "NSynth–OpenL3" configuration, the lowest MAE of 0.030 is achieved with minimal temperature scaling at $T = 1.05$. The reliability diagrams for the ensembles with optimal temperature scaling are added to Figure 1 as blue lines.

CONCLUSION

This study investigates the reliability of confidence values for two tasks in automatic music classification: music genre and instrument family classification. Each of the tasks is approached with two deep neural network architectures—a ResNet trained from scratch and a model using the pre-trained OpenL3 embeddings. In this work, we demonstrated that state-of-the-art deep learning approaches still face limitations in estimating realistic posterior class probabilities for music classification. To tackle this challenge, we explored the use of deep ensembles and temperature scaling, thus enhancing the reliability of probability estimates. It is essential to mention, however, that the optimal value for temperature scaling is both dataset and model dependent, demanding careful selection and adjustment.

In future research, we plan to explore alternative metrics to MAE for assessing the reliability of posterior class probabilities. The current MAE metric incorporates confidences for all classification decisions, including winning and non-winning classes. However, in real-world applications, users are primarily concerned with the confidence of the winning class, as this represents the confidence of the classification decision. Consequently, we propose the utilization of a weighted MAE metric, with weights determined by the frequencies of winning items within each confidence bucket.

Additionally, the reasons for the underconfidence observed for the ResNet trained from scratch on NSynth require further investigations. One possible reason could be the shift between training and test data, which contains different instruments from the same family. Another reason could be the relatively large amount of training data in combination with the selected network architecture.

To conclude, our research contributes to the continued progress in automatic music classification by underlining the significance of reliable classifier outputs and evaluating potential improvements. Integrating these findings into future studies will unquestionably produce more precise, reliable, and valuable music classification systems.

ACKNOWLEDGMENTS

This study was supported by the German Research Foundation (AB 675/2-2) and H2020 EU project AI4Media – A European Excellence Center for Media, Society, and Democracy – under the Grand Agreement 951911.

REFERENCES

[1] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3852–3856.

[2] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A dataset for music analysis. In *Proc. of the 18th International Society for Music Information Retrieval Conference*.

[3] Robert PW Duin and David MJ Tax. 1998. Classifier conditional posterior probabilities. In *Proceedings of the Joint IAPR International Workshops SSPR'98 and SPR'98*. Sydney, Australia, 611–619.

[4] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*. Sydney, Australia, 1068–1077.

[5] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc. of International conference on machine learning (ICML)* (New York, NY, USA). 1050–1059.

[6] Sascha Grollmisch and Estefanía Cano. 2021. Improving semi-supervised learning for audio classification with fixmatch. *Electronics* 10, 15 (2021), 1807.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proc. of International conference on machine learning (ICML)* (Sydney, Australia). 1321–1330.

[8] Yoonchang Han, Jaehun Kim, and Kyogu Lee. 2017. Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 1 (2017), 208–221. https://doi.org/10.1109/TASLP.2016.2632307

[9] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. 2019. 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)*. https://doi.org/10.1109/CVPR.2019.00013

[10] Daniel Kostrzewa, Piotr Kaminski, and Robert Brzeski. 2021. Music Genre Classification: Looking for the Perfect Network. In *Proc. of the 21st International Conference in Computational Science (ICCS)*. 55–67.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[12] Hanna Lukashevich, Sascha Grollmisch, and Jakob Abeßer. 2023. Quantifying Uncertainty in Music Genre Classification. In *Proceedings of The 49th Annual Conference on Acoustics DAGA*. Hamburg, Germany, 1378–1381.

[13] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proc. of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 32.

[14] António Ramires and Xavier Serra. 2019. Data augmentation for instrument classification robust to audio effects. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. Birmingham, United Kingdom.

[15] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2020. Contrastive Learning of General-Purpose Audio Representations. *arXiv preprint arXiv:2010.10915* (2020).

[16] Michael Taenzer, Jakob Abeßer, Stylianos I. Mimilakis, Christof Weiß, Hanna Lukashevich, and Meinard Müller. 2019. Investigating CNN-based Instrument Family Recognition for Western Classical Music Recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. Delft, The Netherlands, 612–619.

[17] Hang Zhao, Chen Zhang, Bilei Zhu, Zejun Ma, and Kejun Zhang. 2022. 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)*. 606–610.