Temperature scaling for reliable uncertainty estimation: Application to automatic music genre classification

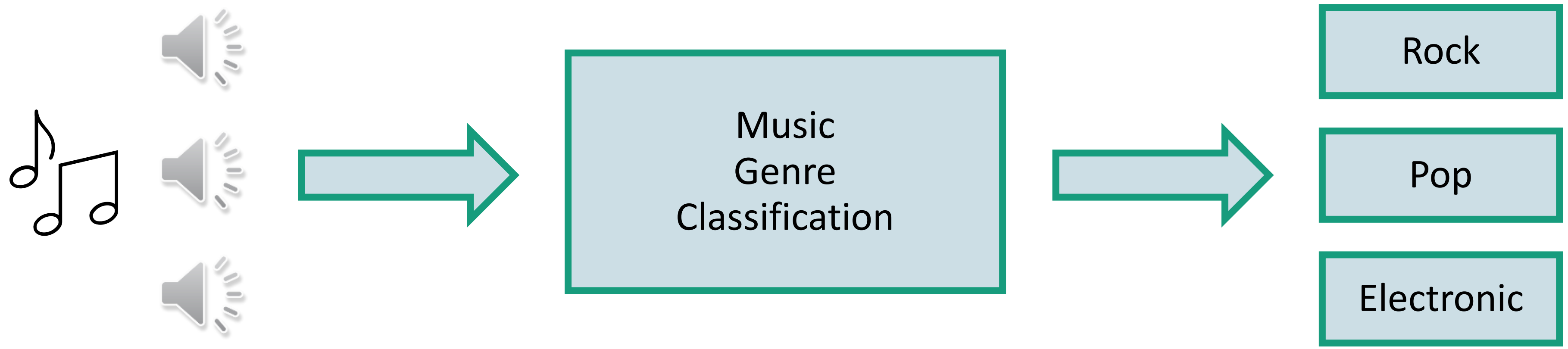
Uncertainty meets explainability @ ECML PKDD 2023

Hanna Lukashevich, Sascha Grollmisch, Jakob Abeßer
Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany

Turin, Italy
20.09.2023

Automatic Music Classification
What is it?

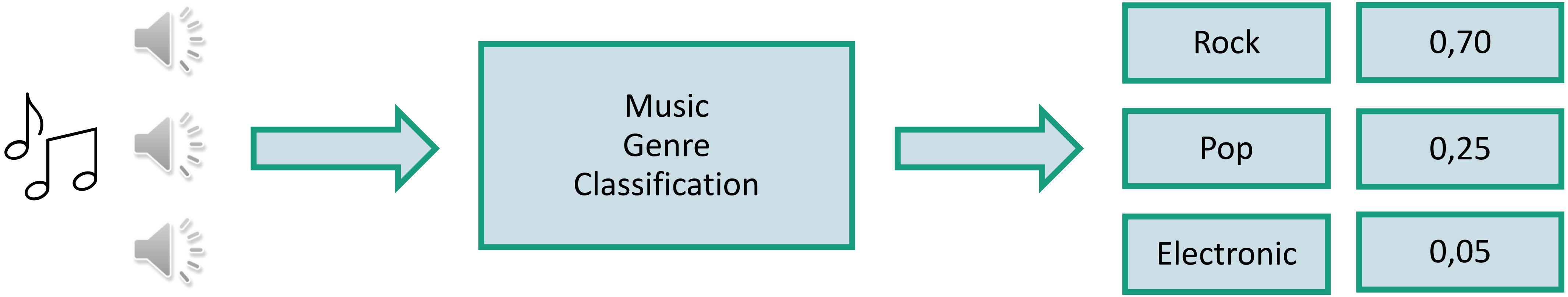
- Music genre classification is crucial for recommendations and content organization
- Common way to describe musical content, also for unknown pieces
- Multiple possible taxonomies, based on the taxonomy of training data
- Quite practical, even while subjective and imperfect



Automatic Music Classification

What is it?

- Music genre classification is crucial for recommendations and content organization
- Common way to describe musical content, also for unknown pieces
- Multiple possible taxonomies, based on the taxonomy of training data
- Quite practical, even while subjective and imperfect
  - Uncertainty of the classifier



Reliable posterior class probabilities

Why is it interesting?

- Why?
  - Deep learning models do not provide 100% accuracy
  - Dealing with uncertainty is a key aspect in real-world applications
  - Humans have a natural cognitive intuition for probabilities
  - Reliable probability estimates can be used to incorporate neural networks into other probabilistic models

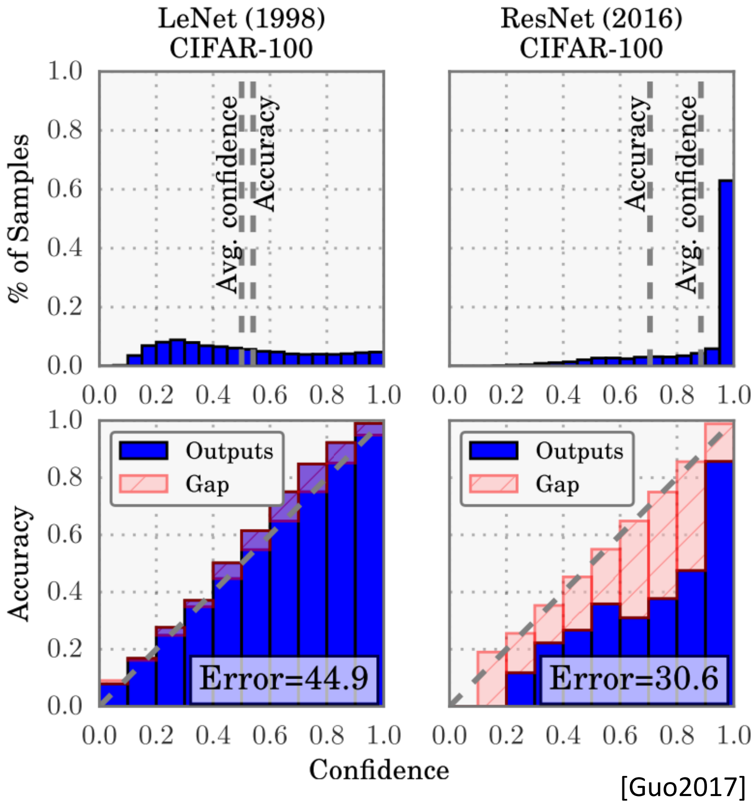
- How does it work now?
  - Ideally: confidence measures as estimations of the probability of correct classification
  - Usually: the estimated posterior class probabilities serve as confidence measures

Fraunhofer IDMT

Deterministic overconfidence

Neural networks are typically overconfident

- Why not using softmax output?
  - Highest softmax output significantly larger than the probability of the corresponding class [Guo2017]
  - **Deterministic overconfidence** is large when
    - data is far away from the decision boundary, out of distribution
    - rectified linear units (ReLU) are used [Hein2019]



[Guo2017]

- Methods to **overcome the overconfidence**
  - Calibrating softmax probabilities, post-hoc, with temperature scaling [Guo2017]
    - useful when data can be considered in distribution
  - Approximating Bayesian inference by MC dropout, Activating dropout during inference [Gal2016]
  - Approximating Bayesian inference with deep ensembles [Lakshminarayanan2017]
    - useful when data is out of distribution [Ovadia2019]

Automatic Music Classification

Dataset and neural architectures

- FMA Dataset For Music Genre Classification: https://github.com/mdeff/fma
  - Large-scale dataset for evaluating several tasks in Music Information Retrieval
  - Small balanced subset: 8,000 30s clips with 1,000 clips per one of 8 root genres
  - Hip-Hop, Electronic, Experimental, Instrumental, Pop, Folk, Rock, International

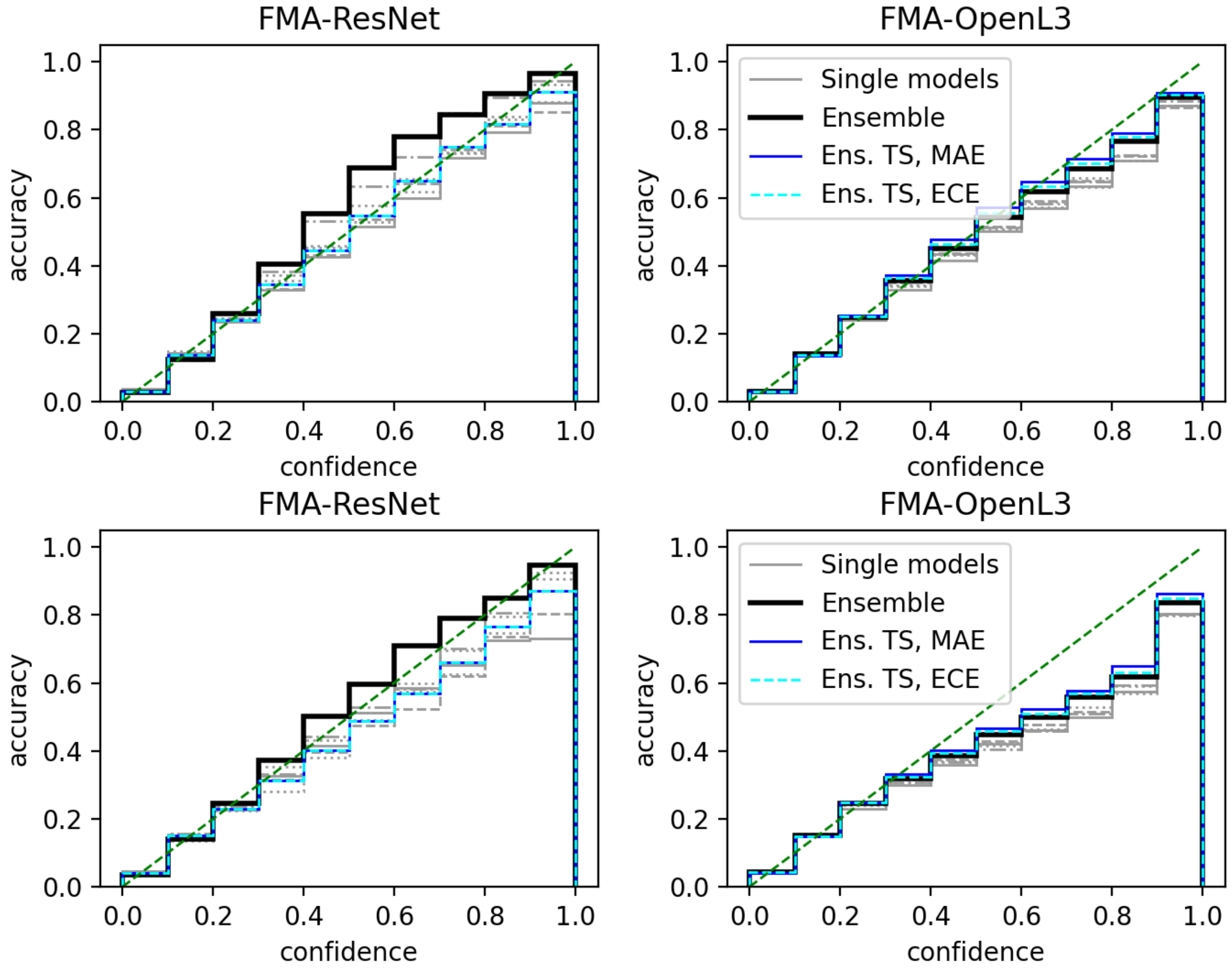
- Two neural architectures: ResNet and OpenL3+MLP
  - ResNet with 420k parameters [Grollmisch 2021]
  - Shallow Multi-Layer Perceptron (MLP) atop pre-trained OpenL3 embeddings [Cramer 2019]

- Evaluation of calibration quality based on reliability diagrams
  - Mean Absolute Error
  - Expected Calibration Error

Reliability diagrams for all datasets and models

Compute the mean absolute error between the reliability curves and the expected accuracy values

- For validation data
- For test data

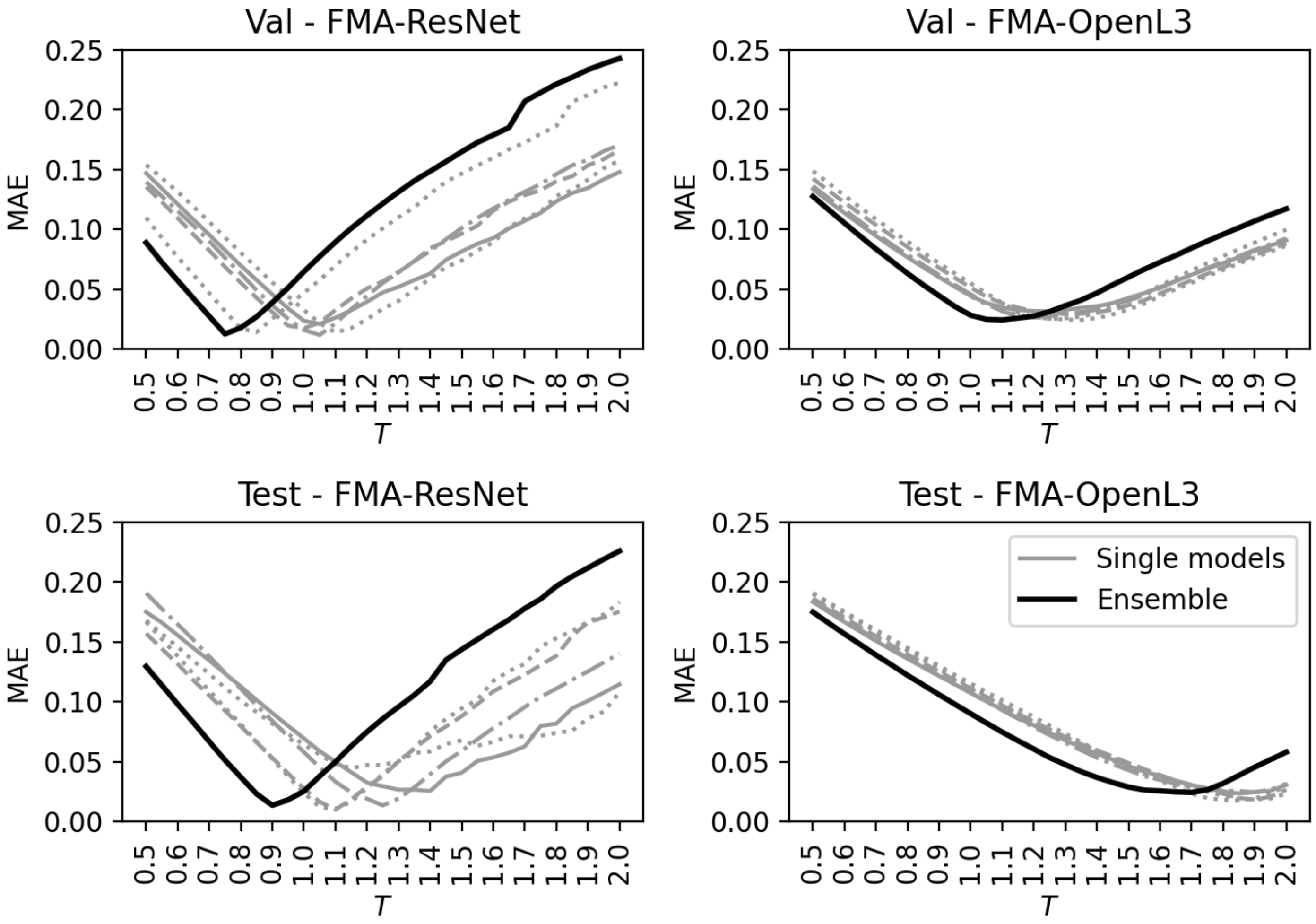


 9/23/2023 Public information

Mean absolute error (MAE) values

Dependency on temperature scaling (T) for all datasets and models

- For validation data
- For test data



 9/23/2023 Public information

Fraunhofer IDMT

Conclusions

Take away message

- Music genre classification is crucial for recommendation systems and content organization
- Neural networks struggle to estimate class probabilities accurately

- Temperature scaling and deep ensembles improve output predictions
- Experiments on the Free Music Archive dataset demonstrate the effectiveness of temperature scaling with deep ensembles
- Various metrics are explored to find optimal calibration temperature

- Discrepancy in optimal temperatures for validation and test data highlights importance of considering generalization capability and data distribution variations

References
Previous work

- [Cramer2019] J. Cramer et al., “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.
- [Gal2016] 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.
- [Grollmisch2021] S. Grollmisch and E. Cano, “Improving semi-supervised learning for audio classification with FixMatch,” Electronics, vol. 10, no. 15, p. 1807, 2021.
- [Guo2017] C. Guo et al., “On calibration of modern neural networks,” in Proc. of International conference on machine learning (ICML), 2017, pp. 1321–1330.
- [Hein2019] M. Hein et al., “Why ReLU networks yield highconfidence 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.
- [Lakshminarayanan2017] B. Lakshminarayanan et al., “Simple and scalable predictive uncertainty estimation using deep ensembles,” Advances in neural information processing systems, vol. 30, 2017.
- [Ovadia2019] Y. Ovadia et al., “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, 2019.

Thank you for your attention!

Fraunhofer
IDMT

Contact

Hanna Lukashevich
Head of Semantic Music Technologies
Tel. +49 3677 467-224
Fax +49 3677 467-467
hanna.lukashevich@idmt.fraunhofer.de