MUSIC GENRE CLASSIFICATION USING 1D AND 2D CNN MODELS

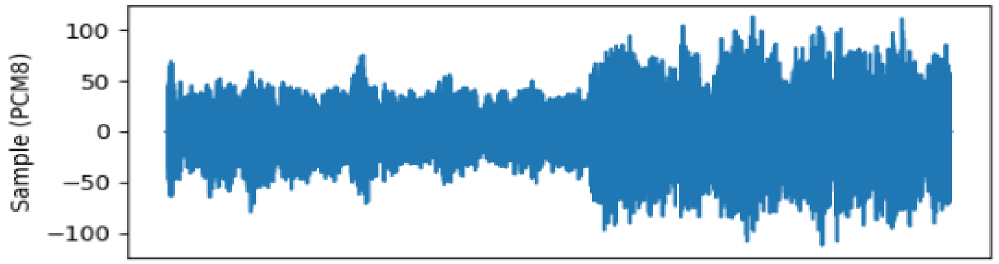
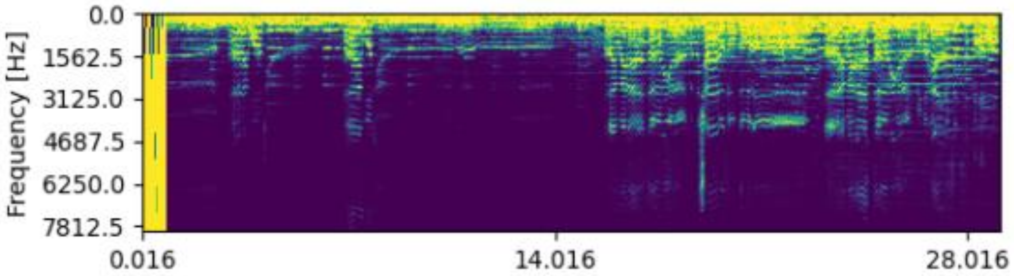
Liviu-Ștefan Neacșu-Miclea
ICA 246/2

SUMMARY

- Music Genre Classification
- Literature Review
- MagnaTagATune Dataset
- Data Analysis
- Proposed Method
- Experimental Results
- Conclusions and Future work

MUSIC GENRE CLASSIFICATION

- Auto music tagging
  - Help search engines keep up with the explosion of media content
- Audio data expressed as

Waveform	Spectrogram
	

- **Question:** Which one can be used to better extract audio features?

LITERATURE REVIEW

- Traditional ML methods
  - K-NN
  - Random Forest
  - SVM
  - Logistic regression
- Deep learning techniques
  - ANN
  - **CNN (1D/2D)**: Musicnn, Harmonic CNN, FCN, CRNN
  - Self-attention

Method	MTAT	
	ROC-AUC	PR-AUC
FCN (Choi et al.)	0.9005	0.4295
FCN (w/ 128 Mel bins)	0.8994	0.4236
Musicnn (Pons et al.)	0.9106	0.4493
Musicnn (w/ 128 Mel bins)	0.9092	0.4546
Sample-level (Lee et al.)	0.9058	0.4422
Sample-level+SE (Kim et al.)	0.9103	0.4520
CRNN (Choi et al.)	0.8722	0.3625
CRNN (w/ 128 Mel bins)	0.8703	0.3601
Self-attention (Won et al.)	0.9077	0.4445
Harmonic CNN (Won et al.)	0.9127	0.4611
Short-chunk CNN	0.9126	0.4590
Short-chunk CNN + Res	0.9129	0.4614

State of the art in genre classification (Won et al.)

MAGNATAGATUNE DATASET

- 25863 audio clips
  - 30 seconds long, 16kHz 32kbps MP3
  - 188 tags
- Songs were labeled by users of the two-player online game platform TagATune
  - A tag was considered valid for a song if more than three users connected that tag to the song.

DATA ANALYSIS (1)

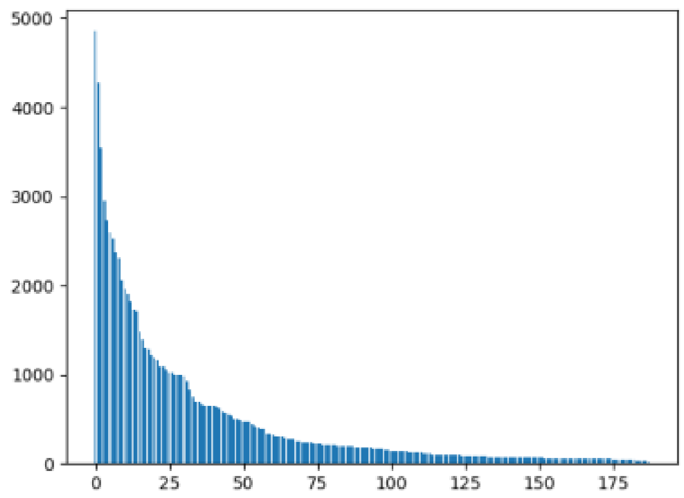


Fig.3.1.1. MTAT labels distributions

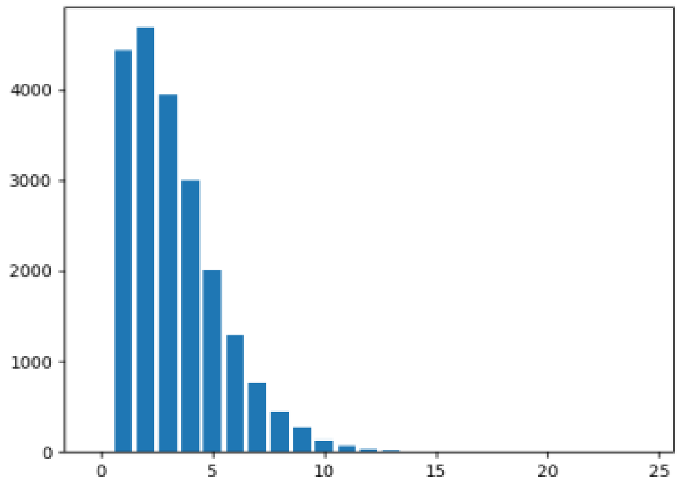


Fig.3.1.2. MTAT instances count per number of tags

- We are facing a multi-class multi-label classification.
- Most literature works perform the top-50 genres classification.

DATA ANALYSIS (2) – TOP50

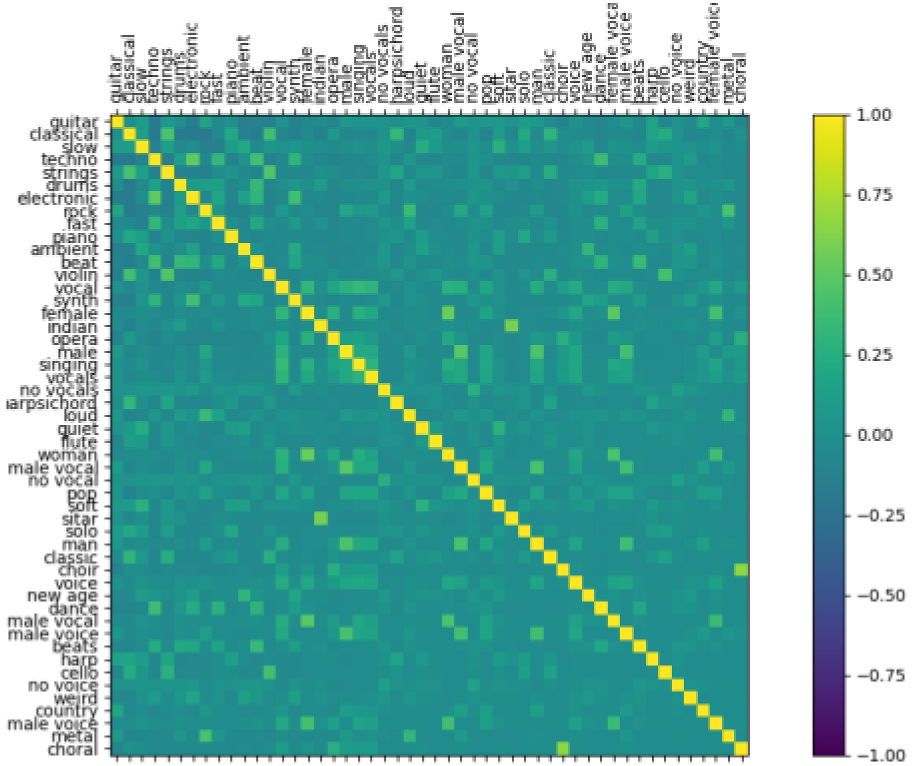


Fig.3.1.3. MTAT top-50 genre correlation matrix

Genre 1	Genre 2	Corr.
choir	choral	0.6573
indian	sitar	0.5910
female	woman	0.5402
electronic	techno	0.5107
female	female vocal	0.4896
male	male vocal	0.4886
strings	violin	0.4535
male	man	0.4403
woman	female vocal	0.4349
metal	rock	0.4263
classical	strings	0.4247
man	male vocal	0.4246
cello	violin	0.4071

Table 3.1.1. Top highly correlated tags

- Top-50 tags may contain redundancies (pairs of tags that have roughly the same meaning and higher correlation)

DATA ANALYSIS (3) – TOP10

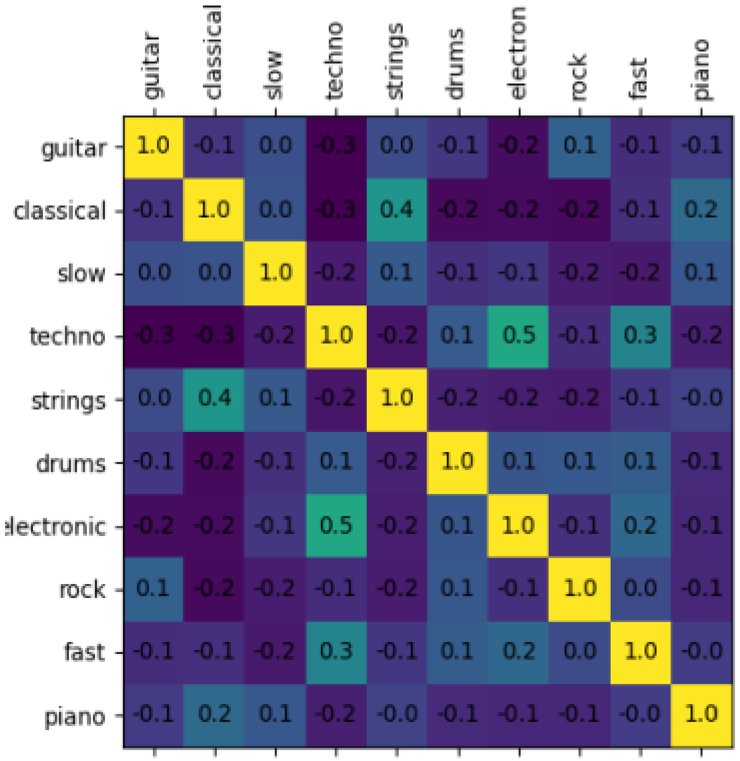


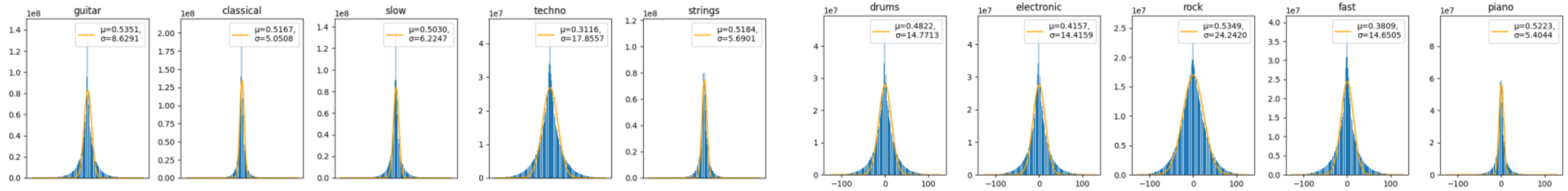
Fig.3.1.4. MTAT top-10 genre correlation

Instances count	**Genre**
4852	guitar
4274	classical
3547	slow
2954	techno
2729	strings
2598	drums
2519	electronic
2371	rock
2306	fast
2056	piano

Table 3.1.2. MTAT top-10

- We therefore shift to top-10 classification
  - At least 2000 samples per label
  - Removed redundancies
  - Less computationally intense
  - Meaningful correlations (classical-strings, techno-electronic)

DATA ANALYSIS (4) – WAVEFORM SAMPLES



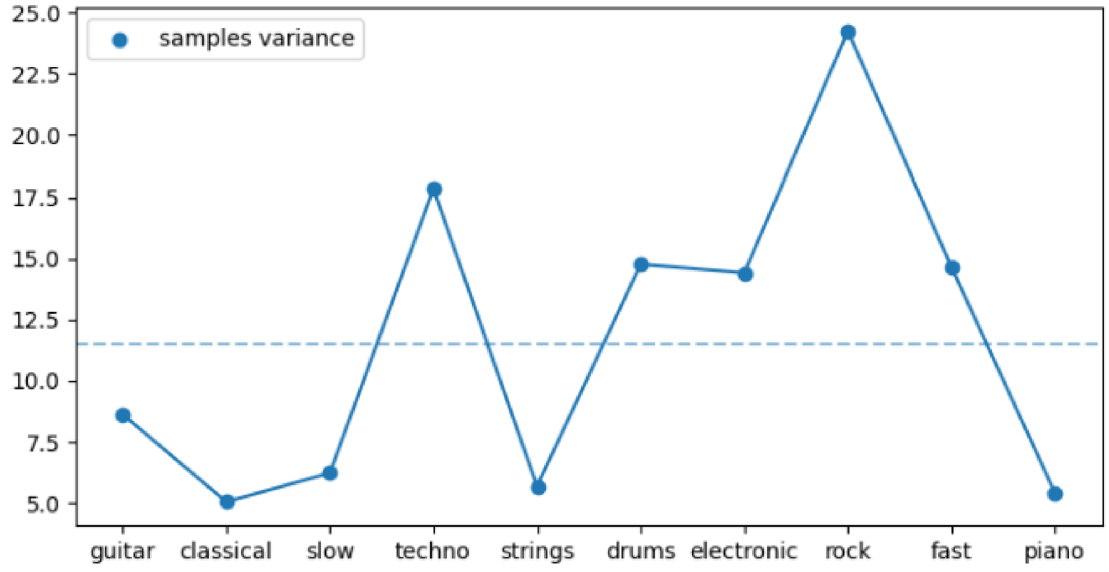


Fig.3.1.6. Sample variance per genre

- The variance of sample densities creates a gap between two types of tags:
  - Low variance: guitar, classical, slow, strings, piano
  - High variance: techno, drums, electronic, rock, fast

DATA ANALYSIS (4) – SPECTROGRAMS

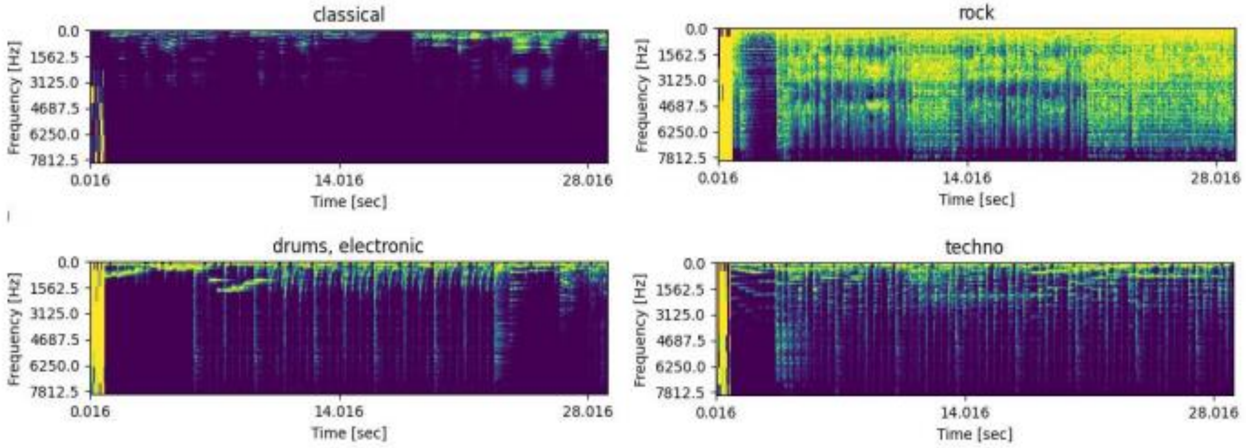


Fig.3.1.7. Spectrogram examples of various genres

- Synthesized music has more dominant higher frequencies
- Sudden drops in frequency distribution
  - Highly paced music or artificial control of the waveform
  - Slow and natural music have a much uniform transition

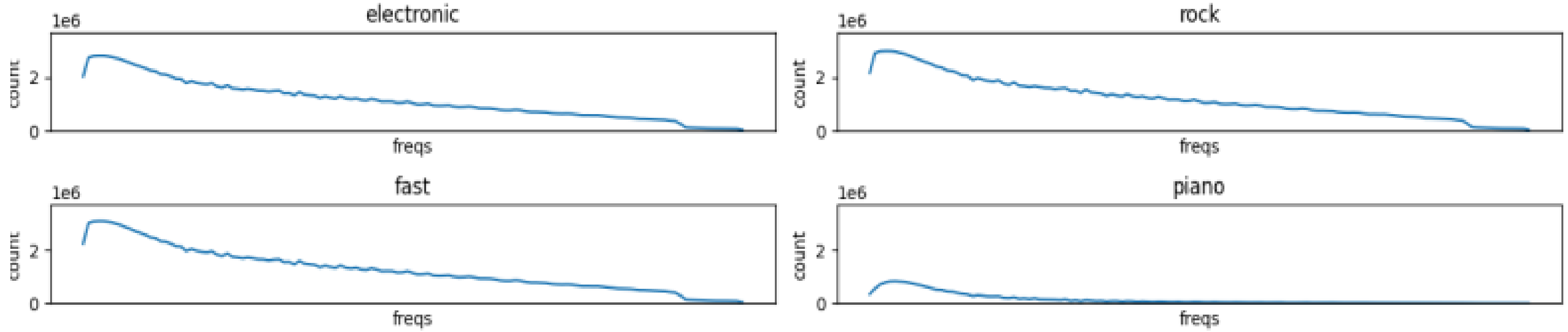


Fig.3.1.8. Time-average frequency distribution per genre

DATA ANALYSIS (4) – DIMENSIONALITY REDUCTION

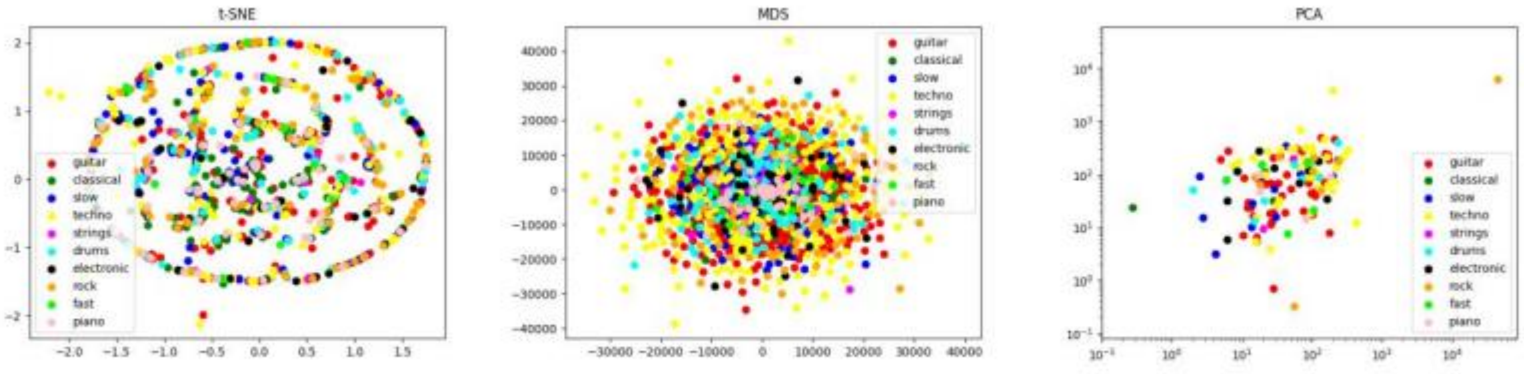


Fig.3.1.9. Sample wave projections

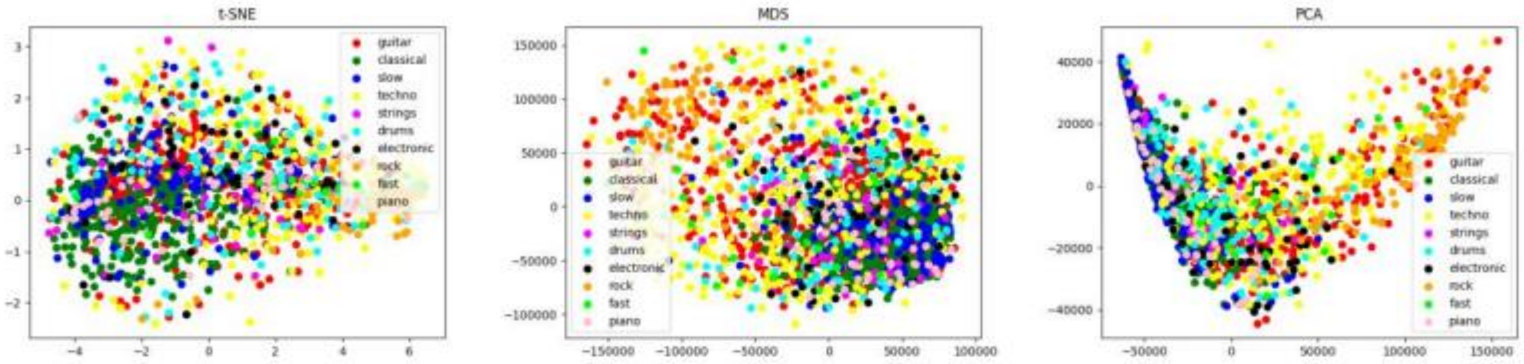
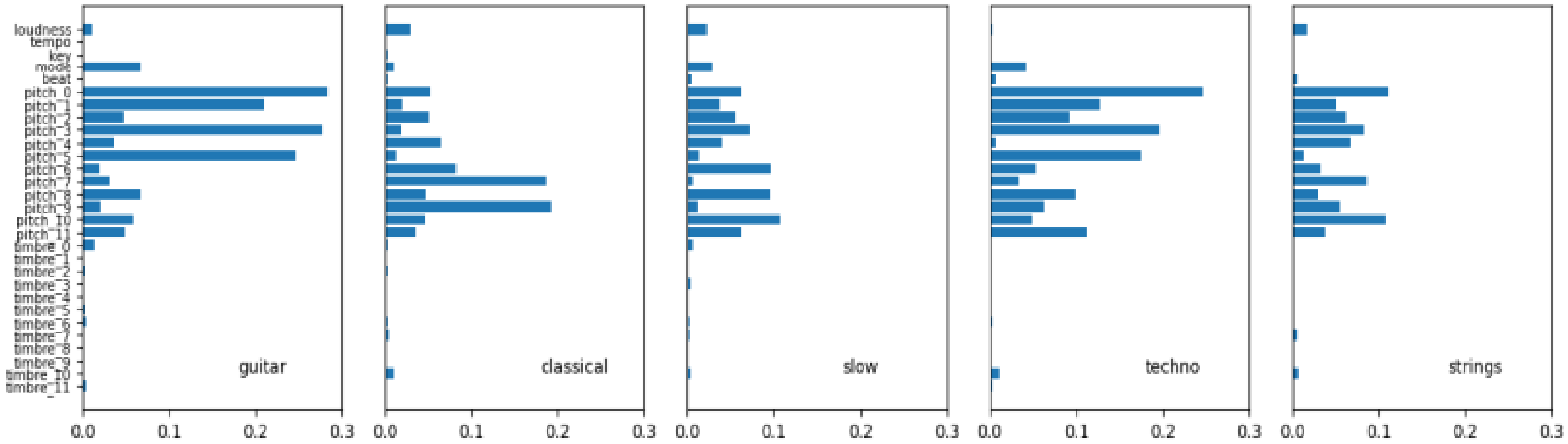


Fig.3.1.10. Spectrogram projections

- Unable to create meaningful DR projections from wave samples
  - Samples are deeply interconnected when related to genre
- Some DR projections (MDS, PCA) of spectrograms reveal the same dichotomy as previous analysis
  - No new enlightments
  - No helpful cluster structure

DATA ANALYSIS (5) – PRE-EXTRACTED FEATURES?

- Music software (like Echo Nest API 1.0) can extract audio features, they come along with MTAT dataset:
  - Loudness, tempo, pitch and timbre vectors
  - Linear regression feature importance reveals the timbre does not matter at all when deciding the 10 tags.
  - Some pitch vectors look more important than others for some tag, but the relations are not crystal clear.



PROPOSED METHOD – EXPERIMENT WORKFLOW

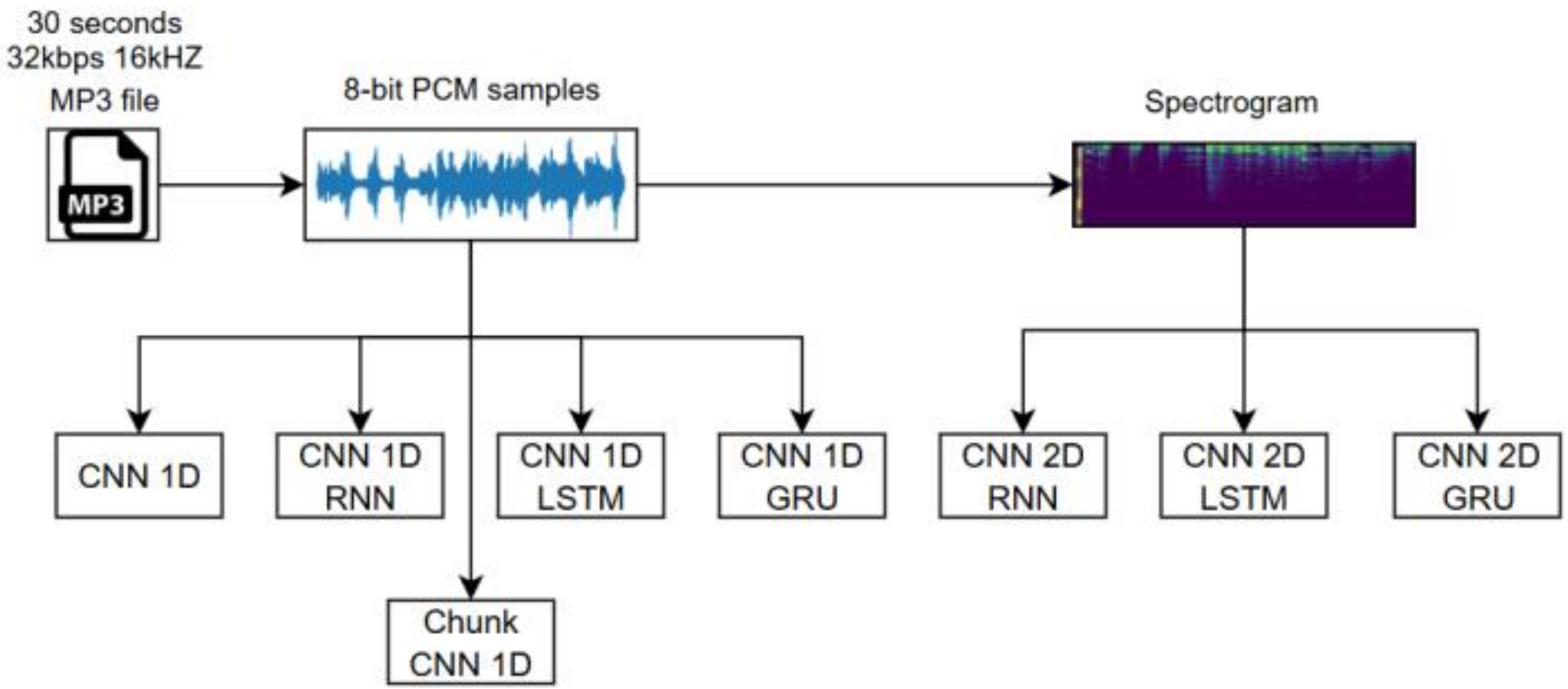
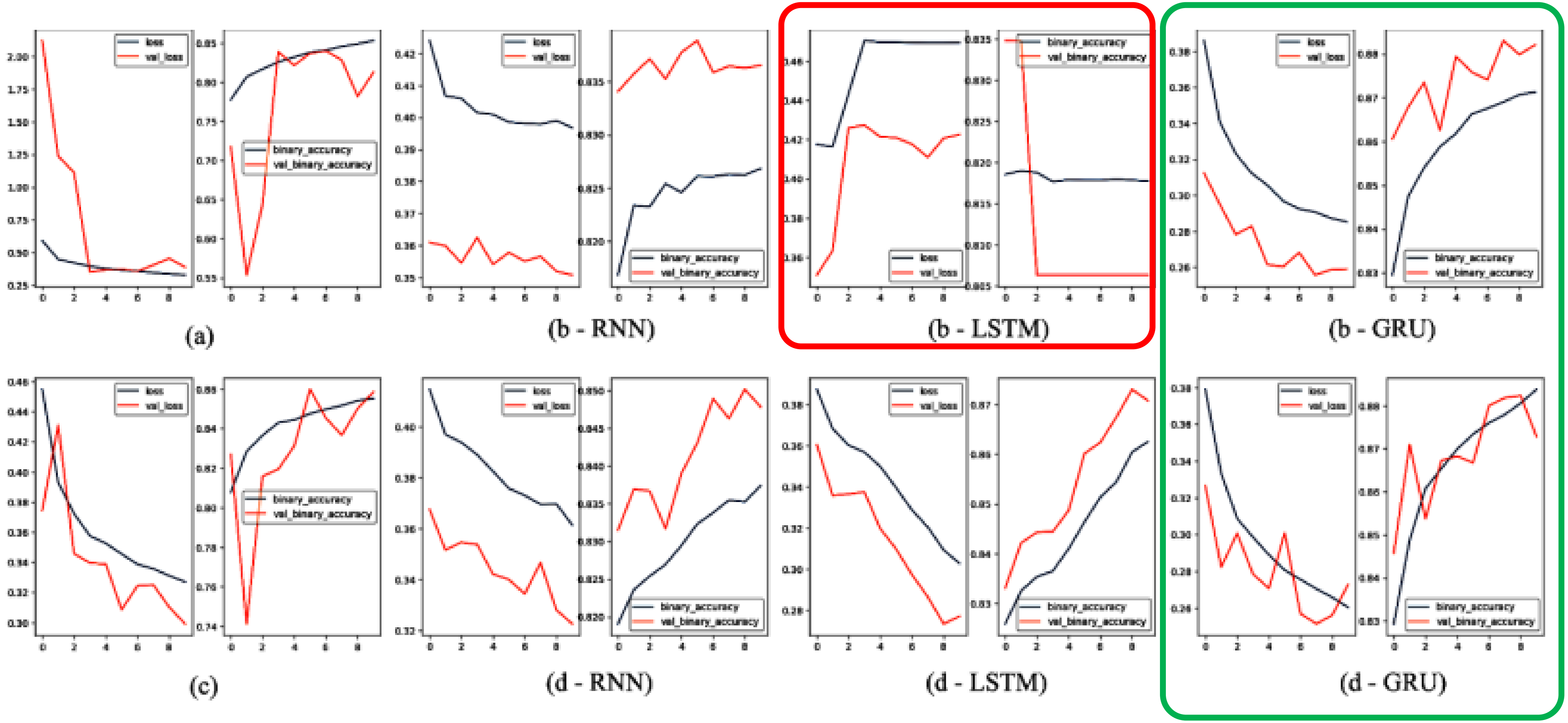


Fig.3.2.1. Experiment workflow

PROPOSED METHOD – CNN ARCHITECTURES

- The usual CNN + Max Pooling [+ BatchNorm] + Dropout stacking
- CNN is connected to Sigmoid activated dense layers
- **Model a:** 1D CNN (on 1D samples)
- **Model b:** 1D CNN + RNN/LSTM/GRU (on 1D samples)
- **Model c:** Chunk based 1D CNN (on 1D samples)
  - (Processes small potions of sound and combines the activations together with a Global Pooling layer)
- **Model d:** 2D CNN + RNN/LSTM/GRU (on 1D samples)
- Training setup:
  - 10 epochs, batch size 16, optimizer Adam, lr 0.001
  - Binary Crossentropy loss, Binary Accuracy metric
  - 9:1 train/val ratio
- Evaluation:
  - Accuracy, Precision, Recall, AUC-ROC, AUC-PR

EXPERIMENTAL RESULTS – TRAINING (1)

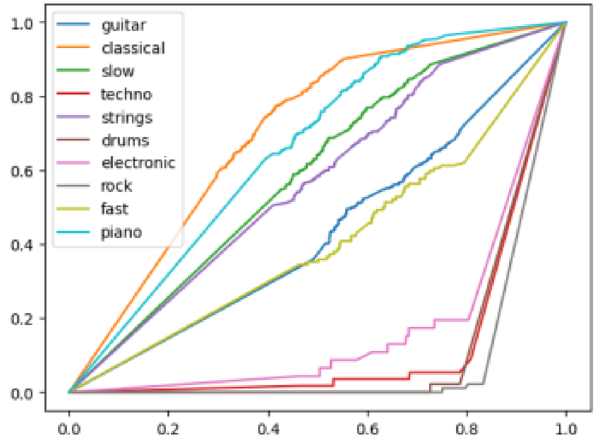


EXPERIMENTAL RESULTS – TRAINING (2)

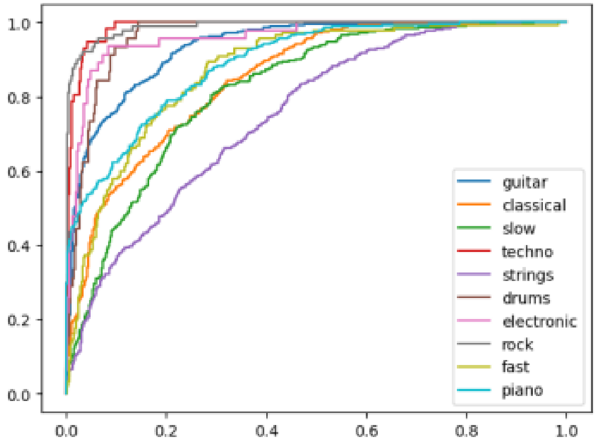
Model	Train loss	Val loss	Train Acc.	Val acc.
(a)	0.3315	0.3888	0.8514	0.8126
(b–RNN)	0.3978	0.3507	0.8262	0.8365
(b–LSTM)	0.4728	0.4226	0.8175	0.8063
(b–GRU)	0.2867	**0.2590**	0.8705	**0.8821**
(c)	0.3311	0.2991	0.8528	0.8584
(d–RNN)	0.3638	0.3227	0.8350	0.8478
(d–LSTM)	0.3062	0.2772	0.8602	0.8708
(d–GRU)	**0.2649**	0.2729	**0.8813**	0.8728

Table 4.1. Train and validation metrics for each model at the end of last epoch

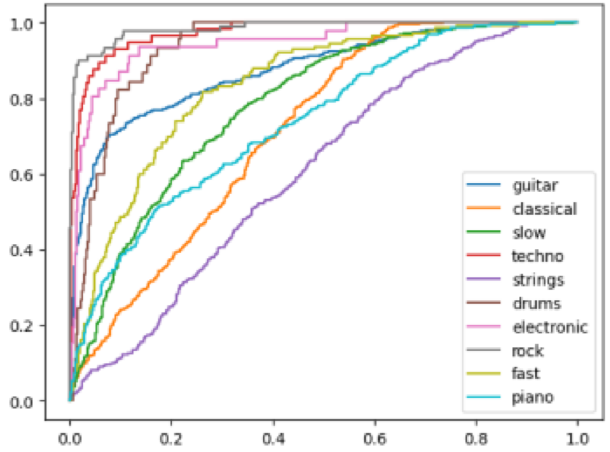
EXPERIMENTAL RESULTS – AREA UNDER CURVES



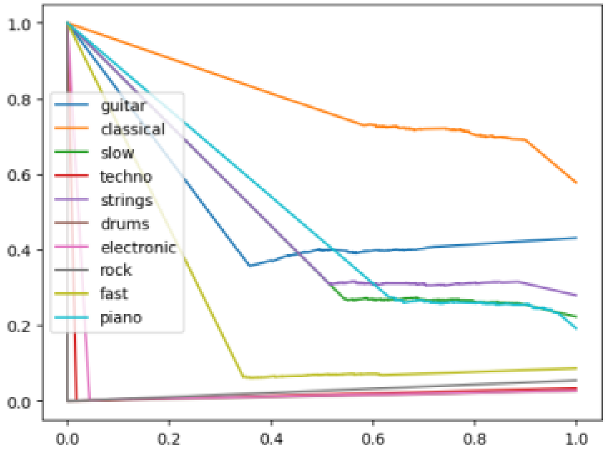
AUC-ROC (b - LSTM)



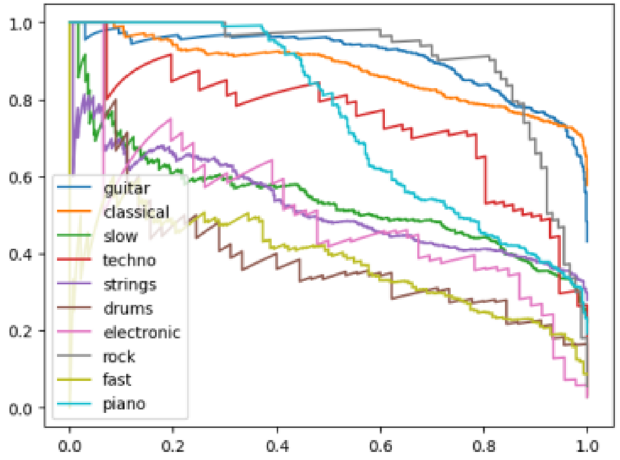
AUC-ROC (d-GRU)



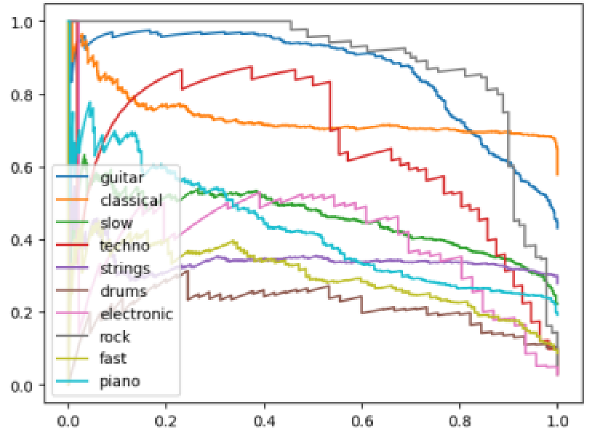
AUC-ROC (a)



AUC-PR (b-LSTM)



AUC-PR (d - GRU)



AUC-PR (a)

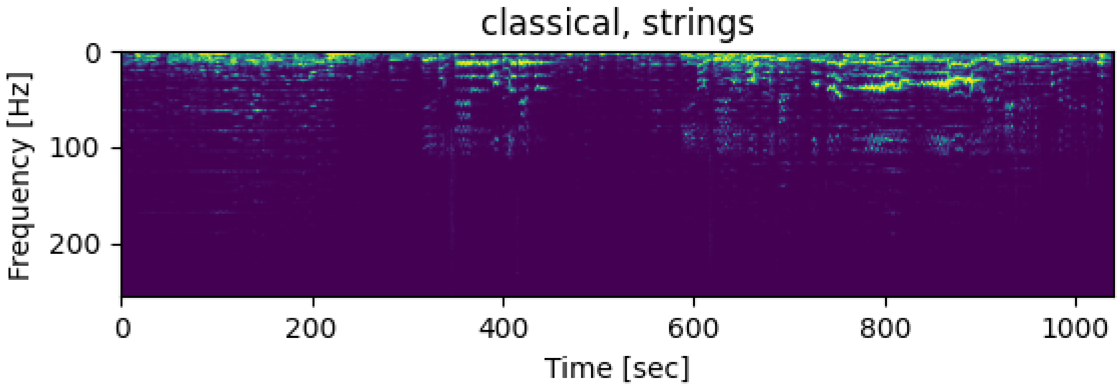
EXPERIMENTAL RESULTS – METRICS

Avg. Metric	a	b-RNN	b-LSTM	b-GRU	c	d-RNN	d-LSTM	d-GRU
Accuracy	0.8125	0.8364	0.8062	**0.8820**	0.8583	0.8477	0.8707	0.8727
Precision	0.8658	0.8551	0.8062	**0.9057**	0.8906	0.8489	0.8888	0.8963
Recall	0.9019	0.9307	**1.0000**	0.9176	0.8855	0.9376	0.9192	0.9176
AUC-ROC	0.8399	0.7222	0.3855	**0.9032**	0.8572	0.8244	0.8735	0.9016
AUC-PR	0.5271	0.3456	0.3079	0.6253	0.5755	0.4990	0.5745	**0.6508**

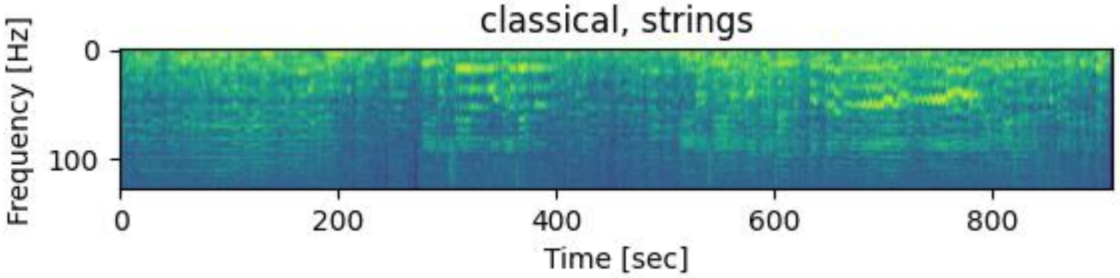
- The failed b-LSTM (CNN 1D) produces only 0 labels.
- Overall, the models have similar performances, leaded by GRU-based CRNNs.

CONCLUSIONS AND FUTURE WORK

- Performance is comparable to SOTA
- Spectrogram-based models tend to perform better, but fairly close to waveform-based models
- Further improvements:
  - 1D CNN over the spectrogram sequence, not just the waveform samples
  - Use Mel spectrograms instead of FFT ones
    - It's said that Mel spectrograms provide a features representation close to what human ear's excitations.



Spectrogram with consecutive Fourier Transforms



Mel spectrogram

THANK YOU FOR YOUR ATTENTION!

BIBLIOGRAPHY

S. Allamy and A. L. Koerich. “1D CNN Architectures for Music Genre Classification”. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (2021), pp. 01–07. url: https://api.semanticscholar.org/CorpusID:234742681.

K. Choi et al. “Transfer learning for music classification and regression tasks”. In: 18th International Society for Music Information Retrieval Conference, ISMIR 2017. International Society for Music Information Retrieval. 2017, pp. 141–149

A. Défossez et al. “High Fidelity Neural Audio Compression”. In: Transactions on Machine Learning Research (2022).

P. Ghosh et al. “A Study on Music Genre Classification using Machine Learning”. In: International Journal of Engineering Business and Social Science 1.04 (2023), pp. 308–320.

D. Kostrzewa, P. Kaminski, and R. Brzeski. “Music genre classification: looking for the perfect network”. In: International Conference on Computational Science. Springer. 2021, pp. 55–67.

E. Law et al. “Evaluation of algorithms using games: The case of music tagging.” In: ISMIR. Citeseer. 2009, pp. 387–392.

M. Matocha and S. Zieliński. “Music genre recognition using convolutional neural networks”. In: Advances in Computer Science Research (2018).

A. V. Oppenheim and R. W. Schafer. Discrete-time signal processing. Pearson, 2010.

J. Stastny, V. Skorpil, and J. Fejfar. “Audio data classification by means of new algorithms”. In: 2013 36th International Conference on Telecommunications and Signal Processing (TSP). 2013, pp. 507–511. doi: 10.1109/TSP.2013.6613984.

M. Won et al. “Evaluation of CNN-based Automatic Music Tagging Models”. In: SMC (2020).