Evaluation of CNN-based Automatic Music Tags Retrieval Models

Team no. 6

1. Chinmayi C. Ramakrishna 181IT113
2. Dolly Gupta 181IT115
3. Shraddha Gole 181IT145

Introduction

- Music tags are provided to characterise music by genres, subgenres, moods, instruments, decades, and languages.
- Music tags make retrieval from large scale music libraries easier.
- Automatic music tagging focuses on the content on the music and aims to predict relevant tags based on acoustic characteristics of the music.
- A tag need not represent the content of the song predominantly. It can represent a chunk of the song.
- Two levels of training are possible for music tagging: song-level training and chunk-level training.

Issues and Challenges

- Music tagging is a multiple instance problem.
- Songs can have multiple features to predict tagging. It is important to find the special specific tags that can recognise a given song.
- The evaluation of content based automatic music tagging system requires consistency in the dataset splits and software versions used for evaluations.

Literature Review

Authors	Methodology	Advantages	Limitations
Automatic Tagging Using Deep Convolutional Neural Networks (2016)	Deep FCN	- Able to capture spatial and temporal dependencies in a spectrogram by applying relevant filters. - Works better than feature extraction using machine learning.	The automatic music tagging is tested on a fixed input length.
End-To-End Learning For Music Audio Tagging At Scale (2018)	Musicnn	Tries to incorporate domain knowledge into its filter designs so that the model can capture temporal patterns.	Uses only chunk level training

Literature Review

Authors	Methodology	Advantages	Limitations
Convolutional Recurrent Neural Networks For Music Classification (2016)	Convolutional recurrent neural network (CRNN)	Uses CNNs for local feature extraction and RNNs for temporal summarisation of the extracted features	CRNN has weak dropout (0.1) between convolutional layers to prevent overfitting of the RNN layers
Sample-Level Deep Convolutional Neural Networks For Music Auto-Tagging Using Raw Waveforms (2017)	Sample-Level Deep Convolutional Neural Networks	- Uses raw audio inputs - Show better results in bigger datasets	There was no analysis on why deep sample-level architecture works well without input normalization

Literature Review

Authors	Methodology	Advantages	Limitations
Toward interpretable music tagging with self-attention (2019)	CNN, Self Attention	- Facilitates the model to learn long-term context by relating each pair of positions directly.	- Adds more weight parameters to the model, which increases training time especially if the input data for the model are long sequences.

Literature Survey

Authors	Methodology	Advantages	Limitations
Data-Driven Harmonic Filters for Audio Representation Learning (2020)	Harmonic CNN	- Utilizes the harmonic structure of the music - shows improvements from previous approaches in terms of ROC-AUC and PR-AUC.	- The model requires more studies and additional investigation on the learnable parameters of the model.

Outcome of Literature Review

- Authors followed different experimental setups (e.g., dataset splits, library versions, different input lengths, computing environments, and optimization methods) when reporting results.
- Different Training methods - chunk level and song-level
- Different datasets were used in different models.

Problem Statement

- Music information retrieval is considered as multi-label binary classification task.
- Most designs are based on convolutional neural networks (CNNs).
- Different papers follow different experimental setups (e.g., dataset splits, library versions, computing environments, and optimization methods) when reporting results.
- Evaluate and compare various Convolution Neural Network based music tagging models under same environmental setting.

Research Objectives

- Use of two training phases namely: song level and chunk level training to handle multiple instance music tagging problem.
- To maintain consistency in evaluation to avoid variation in dataset splits, library versions, computing environments, and optimization methods.
- Assess the robustness against four different types of deformations (pitch shift, time stretch, dynamic range compression, and addition of white noise) and determine their generalization abilities.

Datasets

MagnaTagATune

1. Commonly used datasets for benchmarking automatic music tagging systems.
2. It contains multi-label annotations by genre, mood, and instrumentation for 25,877 audio segments, each 30s long.
3. The audio is in the MP3 format (32 Kbps bitrate and 16 kHz sample rate).

Million Song Dataset

1. In total, this subset of the dataset contains 241,904 annotated song segments.
2. Commonly used for benchmarking on a larger scale.
3. The tags cover genre, instrumentation, moods and decades.
4. The audio segments vary in quality, being encoded as MP3s with a bitrate from 64 to 128 Kbps and sample rate of 22 kHz or 44 kHz.

Datasets

MTG Jamendo

1. Contains audio for 55,701 full songs
2. The minimum duration of each song is 30s, and
3. They are provided in the MP3 (320 Kbps bitrate).
4. The tracks in the dataset are annotated by 692 different tags covering genres, instrumentation, moods and themes.

Methodology

- Train the models in two ways:
  - Train full songs and produce song-level predictions from a song level input.
  - Train the model on short audio chunks and generate chunk-level predictions which is later aggregated in the evaluation phase.
- Implement 4-8 models to predict music tagging.
- Evaluate and compare the models based on appropriate evaluation metrics.

Work Done

- Collection, preprocessing of the dataset which includes extracting Mel spectrograms and re-sampling the audio to 16 kHz sample rate.
- Training the models using a unified optimization method: a mixture of scheduled ADAM and stochastic gradient descent (SGD)
- Implemented 7 models to predict relevant tags for a given music/song.
    - Self Attention
    - Convolutional Recurrent Neural Network
    - Fully Convolutional Network
    - Harmonic CNN
    - Music CNN
    - Sample Level CNN
    - Short Chunk CNN

Work Done

1. **Fully Convolutional Network**

    a. A fully convolutional network (FCN) is a variation of Convolutional Neural Network which consists of only convolutional layers and does not contain any fully connected layers.
    b. There are 4 convolutional layers combined with subsampling layers.
    c. The convolutional layers acts as feature extractor while the fully-connected layer as classifier.
    d. Sampling layers reduces the size of feature map while preserving the information of an activation in the region, rather than information about the whole input signal.
    e. During preprocessing of the dataset, the 29.1s audio segment is converted to Mel spectrogram.

Work Done

2. MusicNN

a. The input is this model is passed in the form of Mel spectrograms.

b. There are vertical and horizontal filters in the first convolutional layer of MusicNN.

c. Vertical filters capture pitch invariant timbral features (harmonic content of the sound wave) present in the mel spectrograms.

d. Horizontal filters are used to capture temporal (time domain) features (like level of energy signals, maximum amplitude) of the audio.

e. During training, the Musicnn model follows chunk level training i.e. it uses only short audio instances of 3s while FCN follows song-level training.

Work Done

3. **Sample Level CNN**

    a. It takes inputs in the form of raw audio waveforms.

    b. This model consists of 10 1D convolutional layers with 1 x 3 filters and 1 x 3 max-pooling layers.

    c. Trained front-end filters replicate the process of generating Mel spectrograms, while back-end convolution layers summarise the results.

    d. Some variations in this model is done by adding squeeze-and-excitation (SE) blocks.

Work Done

4. **Convolutional Recurrent Neural Network**

    a. This model is a combination of CNNs and RNNs.
    b. The local features are extracted by the front end of CNN and then back end of RNN summarizes them temporally.
    c. The front end consists of four convolutional layers with 3 x 3 2D filters and the backend contains two-layer RNNs with GRU.
    d. This model is trained with Long music instances of 29.1s as input.

5. **Self Attention**

    e. It uses CNNs to extract local characteristics and series models for summarizing them.
    f. This model is trained on chunk-level with the 15s-long audio instances.

Work Done

6. **Harmonic CNN**

    a. Harmonic CNN uses harmonically piled time-frequency expression data as input and learnable filters.
    b. There are 128 trainable frequency bands and 6 harmonics for stacking.
    c. Inputs with 5 seconds duration are used for chunk-level training.

7. **Short Chunk CNN**

    d. This model consists of 7 convolutional layers along with one fully-connected layer and has residual blocks (skip connections).

    e. The size of max-pooling layer is 2 x 2.

    f. The audio inputs of this CNN model is of duration 3.69s, hence this model is called short-chunk CNN.

Evaluation Metrics

ROC-AUC

- The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.
- The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Evaluation Metrics

PR-AUC

- It is the Area Under the Precision-Recall Curve to get one number that describes model performance.
- This evaluation metric is useful for a highly imbalanced dataset where the fraction of positive class, which we want to find (like in fraud detection), is small.

Evaluation Metrics

Average Precision

- It is a measure that combines recall and precision for ranked retrieval results.
- The average precision is the mean of the precision scores after each relevant document is retrieved.

$$Average\ Precision = \frac{\sum_r P@r}{R}$$

Evaluation Metrics

Matthews Correlation Coefficient (MCC)

- It is used as a measure of the quality of binary (two-class) classifications.
- It is a more reliable statistical rate that only yields a high score if the prediction performed well in all four confusion matrix categories (true positives (TP) , false negatives (FN), true negatives (TN) , and false positives (FP)), according to the number of positive and negative items in the dataset.

$$N = TN + TP + FN + FP$$

$$S = \frac{TP + FN}{N}$$

$$P = \frac{TP + FP}{N}$$

$$\mathrm{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$$

Results

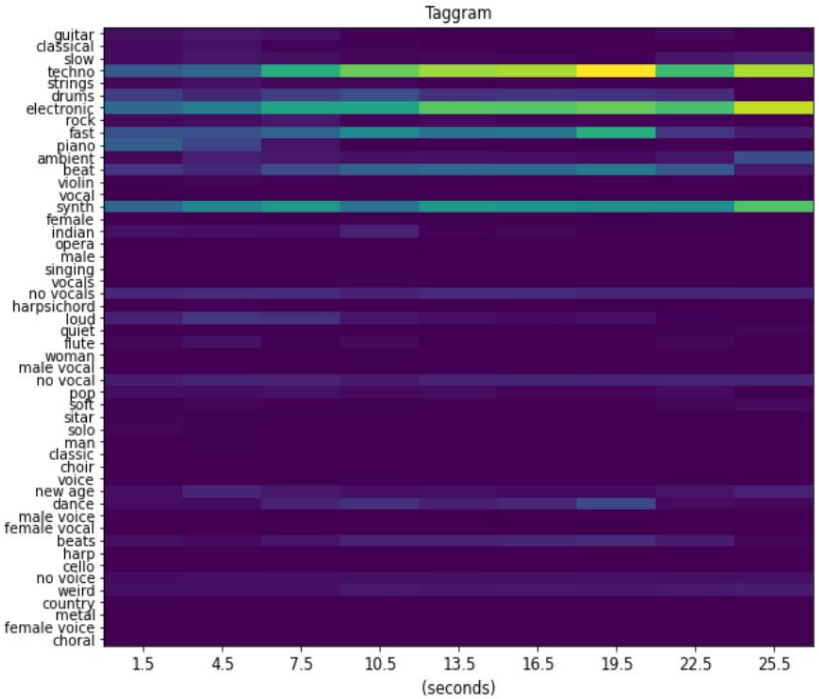

Fig 1. Music CNN Taggram vs time(in seconds)

Results

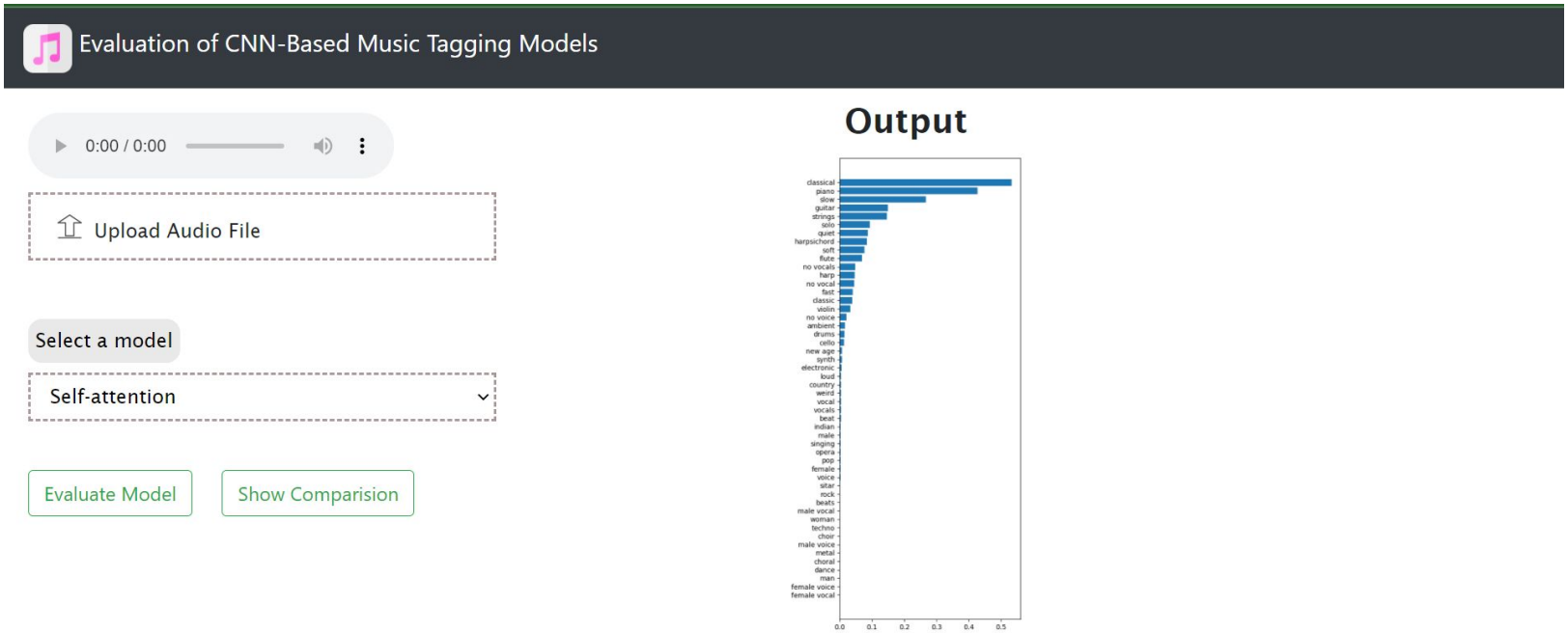

Fig 2. UI 1: To show evaluation of each model

Results

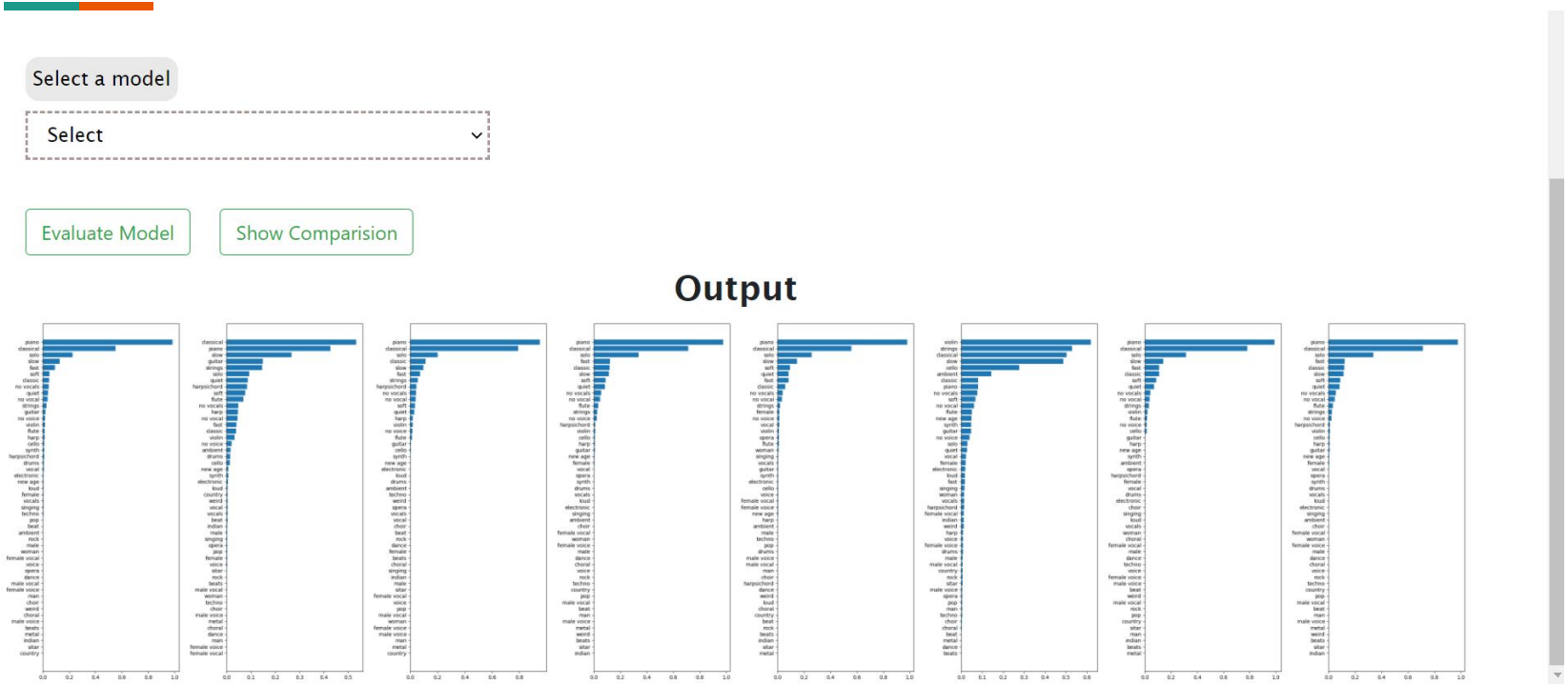

Fig 3. UI 2: Comparison of the models

Results

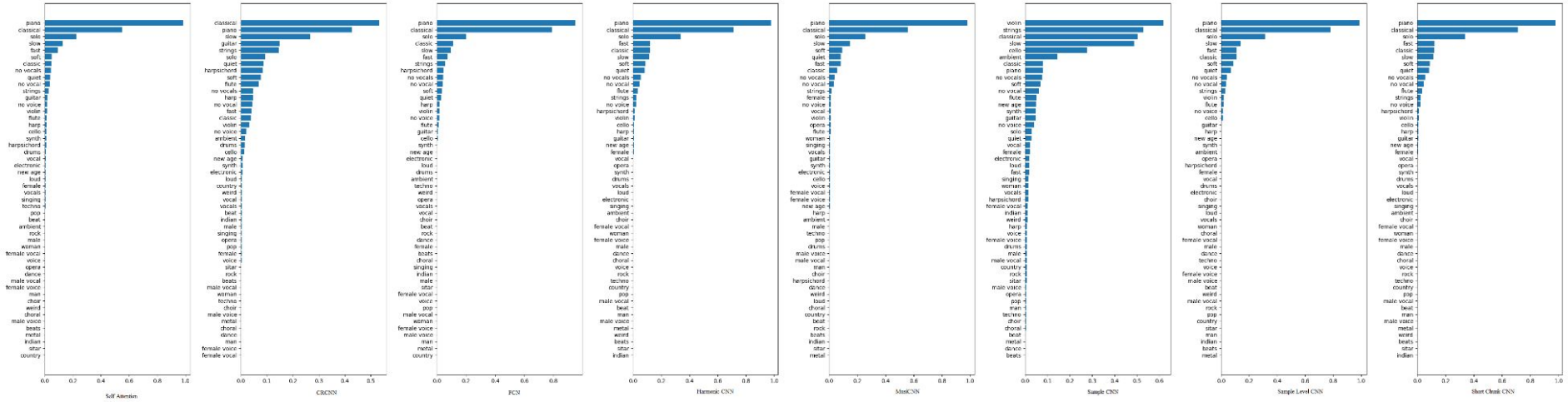

Fig 4. Results for Audio 1

Results

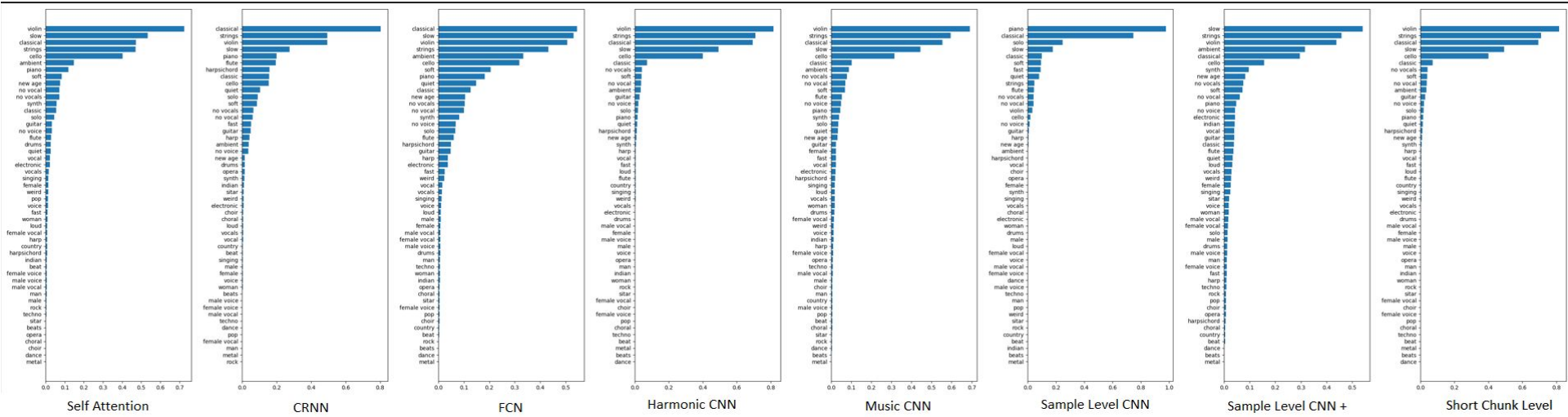

Fig 5. Results for Audio 2

Results

TABLE I: Performance of models on MTAT dataset

Model	ROC-AUC	PR-AUC	AP	MCC
FCN	0.9101	0.4129	0.4221	0.8451
MusicNN	0.9134	0.4519	0.4617	0.8814
Sample-level	0.9103	0.4410	0.4534	0.8510
Sample-level + SE	0.9110	0.4601	0.4790	0.8543
CRNN	0.8899	0.3512	0.3611	0.8371
Self -attention	0.9098	0.4512	0.4600	0.8734
Harmonic CNN	0.9156	0.4611	0.4790	0.8991
Short-chunk CNN	0.9167	0.4799	0.4876	0.8893

Results

TABLE II: Performance of models on MSD dataset

Model	ROC-AUC	PR-AUC	AP	MCC
FCN	0.8712	0.3190	0.3299	0.8350
MusicNN	0.9021	0.2901	0.3030	0.8789
Sample-level	0.8911	0.3010	0.3018	0.8421
Sample-level + SE	0.8923	0.2910	0.3101	0.8400
CRNN	0.8899	0.3112	0.3198	0.8298
Self -attention	0.9098	0.4113	0.4289	0.8683
Harmonic CNN	0.9156	0.4314	0.4521	0.8845
Short-chunk CNN	0.9157	0.4119	0.4478	0.8811

Results

TABLE III: Performance of models on MTG dataset

Model	ROC-AUC	PR-AUC	AP	MCC
FCN	0.8613	0.3098	0.3198	0.8211
MusicNN	0.8521	0.2791	0.2991	0.8701
Sample-level	0.8788	0.2999	0.3001	0.8291
Sample-level + SE	0.8823	0.2888	0.2965	0.8390
CRNN	0.8839	0.2612	0.2790	0.8298
Self -attention	0.9098	0.3213	0.3391	0.8639
Harmonic CNN	0.9156	0.3294	0.3411	0.8789
Short-chunk CNN	0.9167	0.3211	0.3111	0.8778

Individual Contribution

Dolly Gupta (181IT115)	Literature Survey, Self Attention, CRNN, Fully Convolutional Network
Shraddha Gole (181IT145)	Literature Survey, Harmonic CNN, Music CNN, Evaluation Metric 1
Chinmayi C. Ramakrishna (181IT113)	Literature Survey, Sample Level CNN, Short Chunk CNN, Evaluation Metric 2

References

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. of the International Conference on Learning Representations (ICLR), 2015.

K. Choi, G. Fazekas, and M. Sandler, “Automatic tagging using deep convolutional neural networks,” In Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR), 2016.

J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, “End-to-end learning for music audio tagging at scale,” In Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR), 2018.

K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” in Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 2392–2396.

B. McFee, E. J. Humphrey, and J. P. Bello, “A software framework for musical data augmentation.” in Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR), vol. 2015, 2015, pp. 248–254.

THANK YOU