Music Genre and Mood Classification with Convolutional Recurrent Neural Networks

Vincenzo Marciano'
Data Science and Engineering
Polytechnic of Turin
s282004@studenti.polito.it

Vincenzo De Marco
Data Science and Engineering
Polytechnic of Turin
s290373@studenti.polito.it

Michele Di Cataldo
Data Science and Engineering
Polytechnic of Turin
s290091@studenti.polito.it

***Abstract*—In this work it is showcased the adaption of a Convolutional Recurrent Neural Network (CRNN) to two aspects of music classification: Genres and Moods. After pre-processing the audio signals, some audio data augmentation techniques have been explored, to then obtain Mel Spectrograms of the resulting signals. These spectrograms are feeded to the convolutional layers for local features extraction, then a recurrent neural network has been implemented for the temporal summarisation of the extracted features. To train and test the model on genre classification, the GTZAN dataset has been used, while, regarding mood classification, input songs have been acquired from Spotify playlists titled with the corresponding mood. Results obtained have been studied to understand the efficacy of the proposed approach and have been compared with previous works; in particular, the model clearly outperforms the original SVM approach [1] on the GTZAN dataset, showing how relevant neural networks are becoming also for audio signals' classification. Regarding mood recognition, our results show a valuable behaviour across all the songs tested, providing a relevant baseline to take in consideration for future work directions.**

***Index Terms*—convolutional neural networks, recurrent neural networks, genre classification, mood classification**

I. Introduction

Musical genres are labels created and used by humans to describe and better categorize the vast universe of music. Since musical genres arise from a complex view of marketing, historical and cultural factors, there are no strict definitions and boundaries for them; anyway, it is obvious that songs of a particular genre share certain patterns, regarding instruments played, rhythmic structure and pitch content. Moods instead, are way more general traits, and due to the fact that a fair defitinion of mood is *the way a song makes* ***you*** *feel*, they are also much less objective and more personal than genres, making a precise classification more difficult to achieve. Anyway, it is clear that songs sharing the same mood are characterized by some common aspects, ususally related to the song's frequencies and rhythmic structure.

With the growing economic power and importance that online music streaming services (such as Spotify, Apple music, Amazon music, etc.) are gaining, automatic music labeling and classification is becoming more and more essential as the time passes; in fact, since the competition between music streaming companies is growing stronger, a model able to precisely classify music's gernes and moods is extremely valuable for automatic music flow and music recommendation algorithms, and also to better organize the increasingly large numbers of music files these services offer. Therefore, successfully modelling genre and mood classification could yield important benefits for automation in the music industry.

With this in mind, this work tries to adapt the Convolutional Recurrent Neural Network (CRNN) architecture proposed for audio tasks in prior work [2] to the task of genre and mood classification. To this end, the GTZAN dataset has been used for genre classification, while, to obtain input data for mood classification, a method based on the *Spotipy* library has been implemented to get 30 seconds samples from the songs featured in certain Spotify playlists; the playlists have been chosen based on the correspondence between the desired mood and the title of the playlist. Then, a preprocessing pipeline, based on auditory sample extraction through *torchaudio*, has been applied, to have the Mel Spectrograms of input songs as input data: this kind of processed data has proved to be the best choice to gain most of the temporal and frequency features in a packed way, observing an increasing performance boost during the latest years. In particular, the spectrograms represent 3, 5, and 10 seconds splits of the audio files, to evaluate the capability of the model to retrieve relevant information in the fastest yet most efficient way. This data has been feeded to the net, which consists of four sequential convolutional layers, followed by two GRU layers and an output layer: the convolutional part will extract local features from the spectrograms, while the recurrent part will aggregate the temporal patterns of these features.

After assessing the performance of the model on the GTZAN dataset, different audio augmentation techniques have been studied, which will be extensively described in Section III-C. Then, a comparison between the model performance on genre classification and the outcome of the original SVM approach has been carried out in Section IV-B.

Finally, for what concerns the mood classification, the same model is exploited on the dataset created. Then, after minor changes in the Pooling process for a better *feature discrimination*, we observed an enhancement of the resulting model performances.

As a whole, this work makes the following contributions:

1 For what concerns the genre classification, we explored three audio clip lengths, obtaining a model with a very high percentage of accuracy and F1-score.

2 Through mood classification we provide a valuable contribution to a rising auditory environment, since our achieved results are really satisfying, taking in consideration the imbalanceness and plausible inaccuracy of the arranged Spotify dataset.
3 We made possible to visualize the percentage of each tag by using a multi-classification approach, to be able to speculate on how the spectrum of genres or moods is spread along each audio track.

II. Related Works

A. Machine Learning for Audio Processing

Nowadays, research focus has been driven to various fields concerning audio data analysis. Text-to-speech, urban sound event recognition and music suggestions or classification tasks spread out in several Artificial Intelligence departments [3], [4], [5], [6]. In particular, thanks to the improvement gained by exploiting Deep Learning, new approaches and several methods have been explored in music genre classification: the neural network structure, relying on both convolutional and recurrent modules, has proved to outperform previously achieved Machine Learning techniques [2].

Since online sharing music providers, like Spotify, are commonly used nowadays, they are focusing on auditory research to offer a better environment for the user: automatic playlist continuation or daily music suggestions are two of the most relevant research topics [7]. However, it's more difficult to find previous works that generalize the mood comprehension.

B. GTZAN classification baselines

Several studies in genre classification have been made during the years. The first attempt was made by Tsanetakis [1], in which not only the GTZAN dataset but other auditory data was used to exploit an SVM model, fine tuned and runned using features such as MFCC, Spectral Centroid, etc., manually extracted from each audio sample. The gained accuracy was of 61%. The key idea in this work was to pack all the features, processed by a Gaussian model, into single vectors, each describing one of the song used, to then evaluate the model on them.

After various experiments involving the early usage of Deep Learning neural structures, such as multi-DNN [8] or Hybrid Model [9], the state-of-the-art performance is reached by Liu et al. [10], gaining a outperforming 93.9%: they builded a novel BBNN (Bottom-up Broadcast Neural Network), chaining several layer-blocks iteratively.

C. Mood classification baselines

Research studies in mood classification are poorly spread and very different from eachother. The main example comes from a limited study toward Hindi and Western music [11]. Here Patra et al. exploit a FFNN approach, obtaining a 0.751 and 0.835 F-measures, for Hindi and Western respectively. The taxonomy used for this study relies on the Russell's classification [12], in which the individual mood is represented in a 2-dimensional space: for Hindi music 13 different classifications are experienced. This mood taxonomy is extremely useful and it underwent through an important evolution during years, even if we are dealing with two limited (spatially and temporally) families of music pieces.

Another advance in this field was carried out by Weninger et al. [13], where the mood evaluation is determined by a RNN architecture using LSTM. In particular, the goal of this study was to experience on-line mood detection as a regression task, obtaining a 0.704 R^2 score for Arousal mood and 0.500 for the Valence one. Even if good evaluations are achieved, also in this case the NN structure is very simplistic and ignores some spectrograms' features that could add valid informations.

In [14], we start to see some generalization on mood classification, since the work is based on a wider group of songs. Here, two CNNs are run in parallel: the first one extracts the timbral features of each song, while the second one is exploited to retrieve the temporal relationships within the audio spectrum.

A deeply important contribution has given by Choi et al. [2] in exploiting a CRNN structure, aimed to cover all the lacks in performance by previous attempts in music classification, such genre, mood and instruments recognition.

III. Methodology

A. Mel Spectrogram Representation

In the early experiments, several traditional tasks of music classification heavily relied on MFCCs or a subset of audio features. However, in these years there was the urge to shift towards the usage of Mel-Spectrograms: these representations' purpose is to bring with themselves more frequency and temporal contents, making the model achieve better results [2], [6]. Their construction process is as follows:

- Sampling the signal in time domain, accordingly to the sample rate of the input audio
- Decomposing the signal into various Fourier Transforms using Fast Fourier Transforms (FFT), which is a widely used algorithm to compute them efficiently in audio preprocessing. Its expression is showed in the following equation.

$$f[k] = \sum_{j=0}^{N-1} x[j] \left(e^{-2\pi i k/N}\right)^j$$
$$0 \leq k < N$$

- Since it is more likely to find a non-periodic signal as input, Short-Time Fourier Transform (STFT) allows to compute the FFT along overlapping windowed segments of the audio, using the following mathematical expression.

$$STFT\{x(n))\}(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}$$

As a result, we obtain our **spectrogram** in a time-frequency plane.

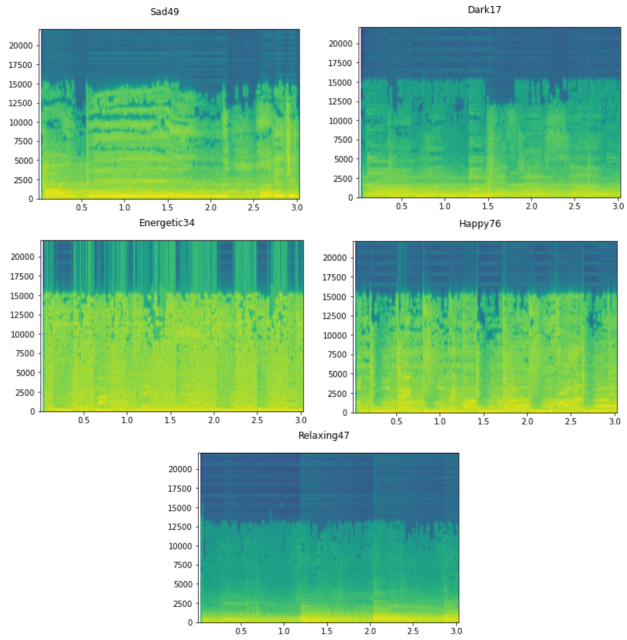


Fig. 1: Spectrograms for all moods of Spotify dataset. Song randomly chosen, excerpts of 3 seconds each.

The spectrograms depicted in Figure 1 and 2 illustrate how temporal and frequency informations are retrieved by their analysis. Spectrograms create visually expressive features: with a certain degree of interpretation, some patterns can be recognized by examining them. These patterns are what we hope to generalize using the proposed NN structure, since they contain the real stylistic value that can be learned by passing a certain input through a Convolutional Recurrent Neural Network (CRNN).

B. Datasets

Music genres for our experiments are retrieved from the ***GTZAN Dataset***. This dataset was one of the early well-defined sets of audio data used for music genre classification in [1]. Performing an exploratory data analysis, we see that the set is divided into ten folders, each one representing a certain genre. Inside every folder, there are one hundred 30-seconds excerpts of different songs in *waw* format. However, during various tests, we deleted a corrupted audio file, *jazz00054.wav*, due to the impossibility to access it during pre-processing.

For mood evaluation, *Spotify APIs*, **Spotipy**, had been exploited: since Million Songs Datasets is quite expensive to use, given its enormous memory requirements, we decided to investigate the top 2021 global playlists shared on Spotify from which we selected the most accurate ones in describing the five moods chosen: *Dark, Energetic, Happy, Relaxing and Sad*. These labels rely on the taxonomy suggested by Russell [12]: for simplicity, the moods belonging to *Valence* and *Arousal* groups will be called v and a respectively, and we will denote with $+$ and $-$ how much a certain mood belong to their class. As Table I suggests, any mood is depicted equivalently belonging to two groups. In reality, some psychology studies

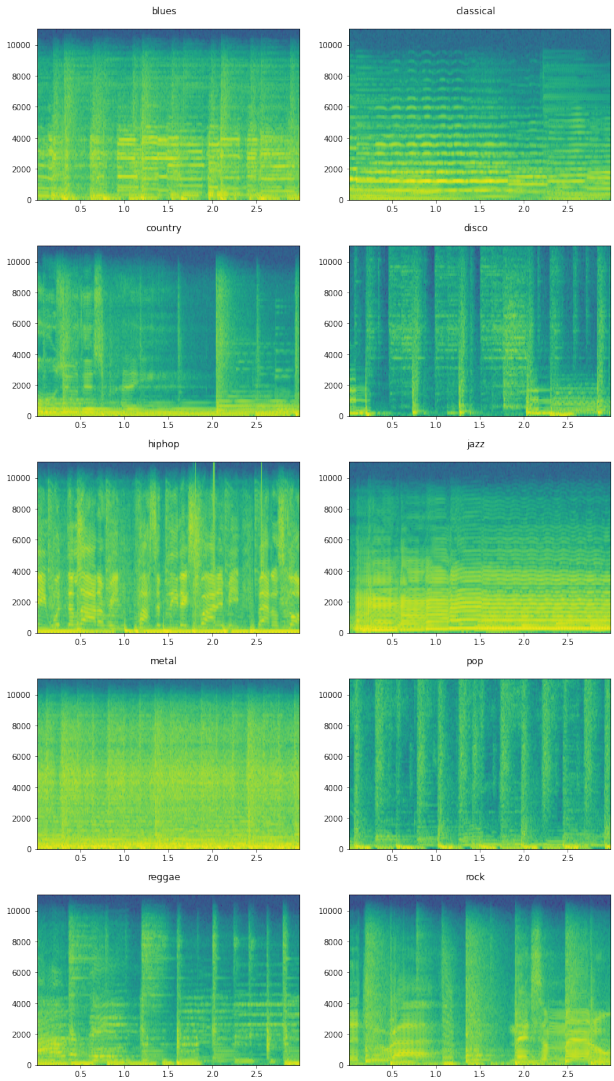


Fig. 2: Spectrograms for all genres of GTZAN dataset. Song randomly chosen, excerpts of 3 seconds each.

Mood	Quadrant
Energetic	*v- a+*
Happy	*v+ a+*
Relaxing	*v+ a-*
Sad	*v- a-*
Dark	*v- a- / v+ a-*

TABLE I: Mood taxonomy by Valence and Arousal attributes

identify a *Dark* mood as a peculiar emotional feeling that is in the middle between the two groups of interest: Figure 2 reflects this distintion. Furthermore, this segmentation is purely subjective, yet globally accepted, since those playlists have actually a large amount of followers, which gives some reliability to the data we are going to use.

The input files are 30s previews of the original songs, stored in *mp3* format. Moreover, the dataset is composed only by the songs that have their preview available for download. Eventually, the dataset could experience some imbalanceness due to the inhomogeneity of the playlists' lenght, this will

worsen the resulting performances of the model.

The training and test sets' size is, as in [6], 90% and 10% of the dataset, then the train set will be split for the training (80%) and validation (20%) processes. Also, from the same study, some standard parameters are retrieved, taking care in adapting them to either GTZAN files with *waw* format, obtained by CD sampling in mono channel, or Spotify's *mp3* files, stereo data with two channels. More details are shown in Table II and III.

Property	Value
Total Number of Tracks	*999*
Total Number of Genres	*10*
Bitrate	*1.4 mbps*
Sample rate	*22 kHz*
Channels	*Mono*

TABLE II: GTZAN parameters for processing

Property	Value
Total Number of Tracks	*326*
Total Number of Moods	*5*
Bitrate	*32kbps*
Sample rate	*44 kHz*
Channels	*Stereo*

TABLE III: Spotify Playlists parameters for processing

C. Audio Processing

In order to create the previously described spectrogram, a STFT is applied on each song. Then the frequency axis is scaled accordingly to *Mel's Log Scale*, followed by a *decibel scale* normalization. The mathematical process is given by:

$$m = 2595 \log 10\Big(1 + \frac{f}{700}\Big)$$

$$d = 10\log_{10}(m/r)$$

The involved parameters are shown in Table IV. This scaling

Property	Value
Number of Mel's bins	*128*
FFT Window Size	*2048*
Hop Lenght	*512*
Reference Power for log-scaling (r)	*1.0*

TABLE IV: Mel's Spectrograms parameters

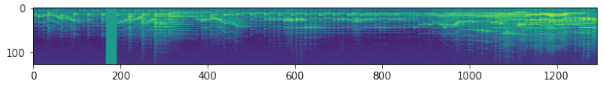
process is nowadays intended as the standard practice for audio preprocessing, as it brings fairly good results in several studies. Moreover, since we are going to deal with several stereo samples for mood evaluation, we have to normalize the two involved channels and consider their average.

For what concerns sample lenght, as in [6], the idea is to feed the algorithm with more training samples, gathered from a certain time window segmentation; in our case, we will try 1s, 3s, 5s and 10s splits. In this way, even if the integrity of the full temporal structure is no longer preserved, the model benefits from obtaining different and shuffled audio samples that mimic a wider independent training set. However, before splitting and shuffling the segments, we should divide the dataset in *train*, *val* and *test* sets, in order to avoid an imbalanced training set.

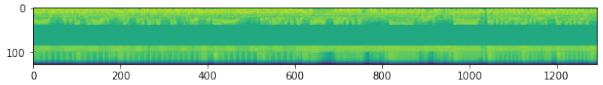
D. Audio Augmentation

Audio augmentation is a standard challenge to deal with to gain a better generalized model, that can achieve good performances. To do so, we rely the entire augmentation process in generating synthetic data through the powerful library *Torchaudio*, which provides several built-in functions that simplify the application process to each song. Particular attention is given to:

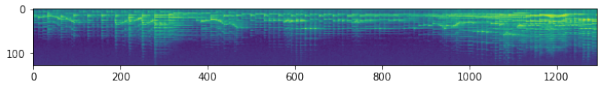
- *Time Masking*: it is applied so that t consecutive time steps [t_0, t_0 + t) are masked, where t is chosen from a uniform distribution, ranging from 0 to the time mask parameter T, and t_0 is chosen from [0, τ - t). The purpose is to emulate a time interval of missing audio, like damaged CDs in car stereos. We set T=50;



- *Frequency Masking*: it is applied so that f consecutive mel frequency channels [f_0, f_0 + f] are masked, where f is chosen from a uniform distribution, ranging from 0 to the frequency mask parameter F, and f_0 is chosen from [0, υ - f). υ is the number of mel frequency channels. It works like a bandpass, lowpass or highpass filter by masking the chosen set of frequencies and detecting the remaining ones as the main components. We set F=60.



- *Time Warp*: Given a log mel spectrogram with τ time steps, we view it as an image where the time axis is horizontal and the frequency axis is vertical. A random point along the horizontal line, passing through the center of the image, within the time steps (W, τ - W), is warped along that line, either to the left or right, by a distance w, chosen from a uniform distribution ranging from 0 to the time warp parameter W. The idea is to shift the spectrogram in time by using interpolation techniques to squeeze and stretch the data in a randomly chosen direction. We set W=20.



- *Adding Background Noise*: we inject some random noise by exploiting Signal-to-Noise-Ratio (SNR), in order to experience the real word imperfections in recording some raw audio sample.

E. Model Structure

Since their early success, Convolutional Neural Networks (CNNs) saw a wide spreading of usage all over the most recent studies, in particular for what concerns image classification. Also in music recognition, these deep learning networks seem to be the best choice to collect and extract the lower layers' features, which then will be processed.

The same fame in audio classification is achieved using Recurrent Neural Networks (RNNs), mostly in speech detection, in which they are used to summarize the temporal structure of the CNNs' outputs.

As a combination of these two contributions, **CRNNs** outperforms the other techniques in various auditory data classification. An important example is based on [6]: in this study, the same CRNN used in [2] is exploited in artist recognition.

As stated above, two types of experiments are carried out: one for genre classification and another for mood recognition. The architecture used in each case is almost the same; the only difference relies on the Pooling layers. A 2D *MaxPoling* is adopted in [6], and our experiments confirmed that it is an excellent choice for the genre recognition task, since each genre spectrogram experiences a variational channel value, in which the maximum value determines the label we are trying to infer. However, for what concerns mood evaluation, the separation between each label is not as evident as in genres, as Figure 1 suggests. For this reason, we switch to a 2D *AvgPooling*.

The number of layers, batch normalization, kernel sizes, filters and padding are adapted from [6], as long as the recurrent unit, in which we expolit two stacked GRUs instead of LSTM, due to their lower number of tunable parameters and better performance in feature retrieval. Since the backend environment we are referring to is PyTorch, we avoid the usage of *SoftMax* activation to work properly with the *CrossEntropy* loss function, which accepts also negative values. As optimizer, after testing *SGD*, we confirmed that *Adam*, as used in [6], is the best choice in terms of performances. We also developed an *Early Stopping* feature with two parameters: *Delta*, the required improvement of the loss to consider a new minimum, and *Patience*, the number of epochs without loss improvement before stopping the training process. Learning Rate as well as Early Stopping parameters are taken into account, trying different parameters' configurations to obtain the best performance.

IV. Results and Discussion

A. Achieved results

1) GTZAN: The scores obtained using different slicing windows are shown in Table V. It's evident that the **3s** window achieves an excellent performance, and this configuration also provides the most stable training loss with respect to the other experiments carried out: Figure 3 depict the loss trend during the training process. However, if the 1s window still achieves fairly good results, the longer segmentations manifest a drop of accuracy and F1 score; this is caused by the lack of generalization resulting from excerpts with too much information to process.

The reported results have been obtained using the parameters' configuration described in Table VI.

Considering the Mel Spectrogram-based architectures, and taking as a comparison the state-of-the-art results described in [10], our model would be outperformed only by the recent yet more complex BBNN structure.

Time Slicing Excerpts	Accuracy	F1-Score
1s	*0.890*	*0.889*
3s	***0.920***	***0.922***
5s	*0.690*	*0.679*
10s	*0.570*	*0.534*

TABLE V: Accuracy achieved for each proposed song slicing window

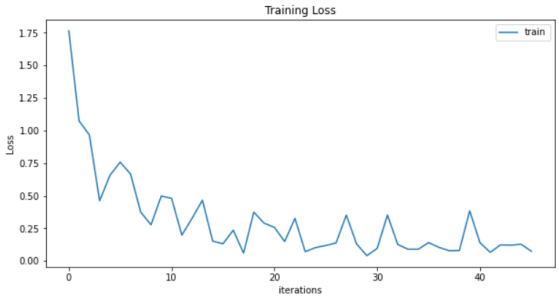


Fig. 3: Training Loss behaviour for the 3s slicing window in Genre Classification.

Parameter	Value
Batch Size	*16*
Learning Rate	*0.001*
Dropout	*0.3*
Patience	*15*
Delta	*0.01*

TABLE VI: Hyperparameters Settings

As a final observation, the GTZAN dataset covers only a few of the entire family of genres, which is continuously evolving, and some of the 10 genres covered are also very similar between eachother (eg. jazz and blues). We could suppose that the proposed genres are not enough to describe the actual musical background, and this lack of data will negatively afflict the generalization of our model.

2) Mood: To get Mood Classification performances each song has been divided into 3s excerpts, since the previous experiments confirmed it as the best window's lenght choice. The model achieved **0.74** *Accuracy* and **0.74** *F1-Score*. The training loss trend is shown in Figure 4.

Also in this task, the parameters' configuration are the ones described in Table VI.

Even if this score could seem low for a classification task, it is important to consider that properly evaluating emotions from an algorithmic point of view is very difficult. While reviewing papers one can notice that the vast majority of approaches assume one mood per track, instead of choosing a multi-classification-based approach; in addition to that, many studies use their own model and taxonomy, hence this makes difficult for us to find the proper straightforward comparison.

Another point we wish to emphasize is the low number of songs used and the imbalanceness of which our dataset suffers, that may result in worse results, along with the software's sound normalization that cuts from a certain level of frequency (as it can be seen on Figure 1).

Although we may consider these settings as a sort of

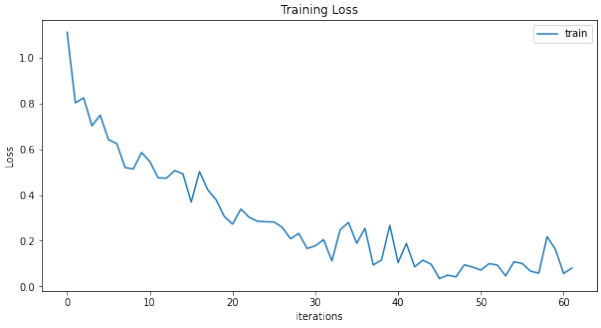


Fig. 4: Training Loss behaviour for the 3s slicing window in Mood Classification.

drawbacks of the proposed task, the model behaves very well with the unseen songs for which it was fed.

3) Data Augmentation: Table VII summarize the results achieved by applying the proposed data augmentation techniques on the model trained on the GTZAN dataset, with a 3s slicing window. We observe that our model performs well and has similar results with time masking, while the performances are a little worse with frequency masking, time warping and background noise.

Augmentation	**Accuracy**
Time masking	*0.89*
Frequency masking	*0.78*
Time warping	*0.81*
Background noise	*0.82*

TABLE VII: Accuracy achieved for each proposed augmentation technique

B. CRNN vs. SVM approach

Genre classification has as starting point the performances of the SVM approach developed in [1]; we replicated their work to better understand the SVM results.

The structure relies on a Support Vector classifier, in which the hyperparameters are tuned by a *CalibratedClassifier*. In this case, we are not dealing with Spectrograms of any type: other kind of audio features are explored, relying on *Librosa* feature extraction functions. In particular:

- *Spectral Centroid*, defined as the center of gravity of the magnitude of STFT spectrum;
- *Spectral Rolloff*, defined as the frequency R_t belowe which 85% of the magnitude distribution is concentrated;
- *Spectral Flux*, defined as the squared difference between the normalized magnitudes of successive spectral distributions;
- *Time Domain Zero Crossings*, provides a measure of the noisiness of the signal;
- *MFCCs*, maybe one of most important feature to deal with in audio processing. After applying Mel-scaling on the computed STFT, we decorrelate the feature vector through a discrete cosine transform.

SVM is the very first idea from which the entire deep learning environment was born. Moreover, considering its capability in understanding non-linear spaces, Support Vector classifiers are quite suitable for pattern recognition.

Anyways, NNs can be considered as the evolution of SVMs, in fact in our specific case study a well-tuned CRNN clearly outperforms the SVM results. Also, we extract deep features directly from Mel's spectrograms (a 3D tensor) instead of working with a single vector of manually extracted features. In addiction to that, we can also suppose that the previously described audio features are not enough to recognize a specific genre since we are missing any temporal informations, while, using a RNN module in our model, we retrieve from each Spectrogram its temporal structure and we use it to summarize the CNN output features.

This allows us to enhance the accuracy gained by the SVM by 50% in the genre classification task.

C. Behavior analysis with new songs

Several songs was passed to our algorithm. However, we focus to examinate three media from three very different, yet very experimental artists: *Alt-J*, *Bon Iver* and *Moderat*. The purpose of these choices was to understand how far the level of the model generalization could go. Moreover, for simplicity and sake of visualization, only the top 3 genres and moods for each song are displayed, while the remaining ones are grouped in a unique label "Other".

The audio tracks proposed are:

- 715-CRΣΣKS by Bon Iver;
- Fitzpleasure by Alt-J;
- Reminder by Moderat.

In the end, results are quite promising and accurate both for mood and for genre, taking into account the subjective point of view of the evaluation. The Figure 5 and 6 confirms our previous assumption: each song may experience an intensive extension of the emotions and genres that it suggests to the auditor. For example, *Reminder* seems to be a Sad song because of the presence of minor chords and owling-mimic vocal modulations, while keeping a strong Energetic component; an interesting evaluation is the one referred to *Fitzpleasure* genre prediction, where we see Rock and Pop as mainly instances, mixed up with Metal due to the presence of "grungy" sounds that evolve along all the song. Last but not least, *715*'s predictions perfectly reflect its Sad tones and Dark atmosphere, assisted by the presence of Jazz chords modulation and Pop influences.

V. Conclusions and Future works

We proposed a Convolutional Recurrent Neural Network and we adapted it for genre classification, on the GTZAN dataset, and mood recognition, using a dataset created with relevant Spotify playlists, retrieving songs' previews by exploiting *Spotipy APIs*. Even without an appropriate hyperparameters tuning, possible future implementation, we tried different configurations, starting from the parameters used by [6] in artist recognition, and found the performances obtained satisfying, considering the drawbacks explained in Section IV.

Possible future directions are:

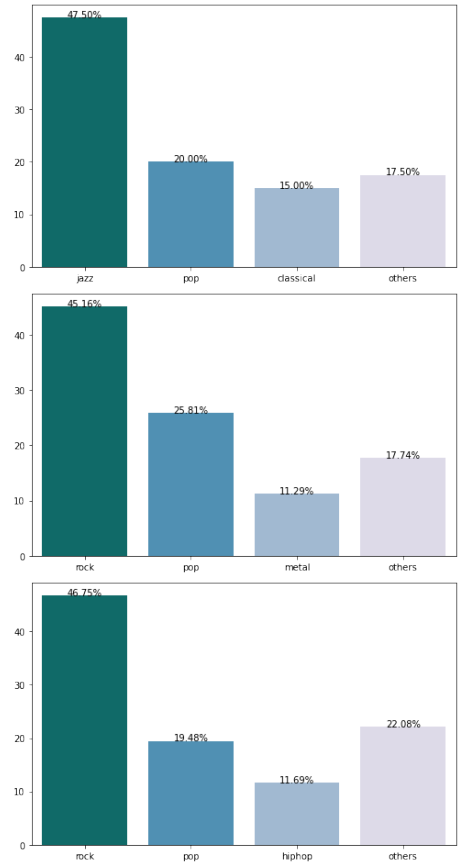


Fig. 5: Genre Prediction of the three songs. The multi-classification shows how the various genres are spread into each song.

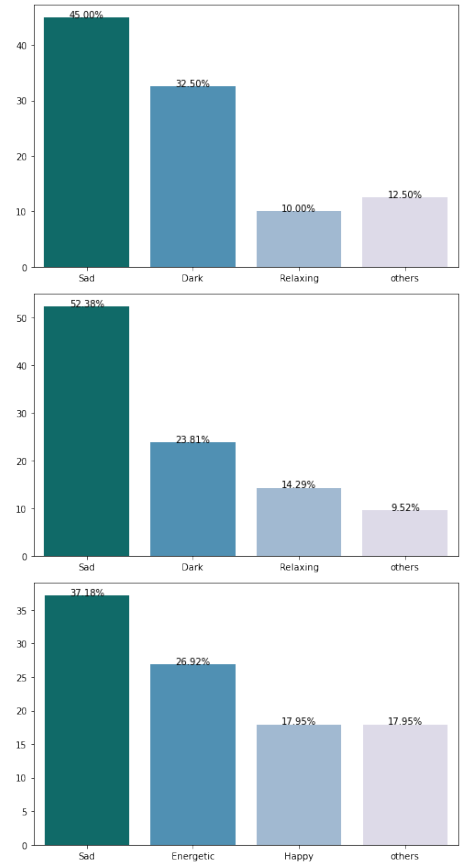


Fig. 6: Mood Prediction of the three songs. The multi-classification shows how the various genres are spread into each song.

- *Parallel NNs*: Parallel deep Neural Networks could be implemented to enhance the performances; a promising example could be exploiting sentiment analysis through speech recognition on the songs' lyrics;
- *Million Song Dataset*: Using as input data this huge dataset could certainly improve performances and generalization capabilities of the model, since it would be able to work with many more genres and moods than in our work; in fact, using Mel Spectrograms and a deep CRNN, unlike in [1], we could discover new patterns and produce more outcomes to study;
- *New activation functions*: Even if ELUs activation funtions have experienced a great success in music tagging, PELU (Parametric Exonential Linear Unit) [15] could be a better choice, since it iteratively adapts values discrimination by using a learnable parameter accordingly to each sample.

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] K. Choi, G. Fazekas, and M. Sandler, "Explaining deep convolutional neural networks on music classification," *arXiv preprint arXiv:1607.02444*, 2016.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and trends in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] Z. Nasrullah and Y. Zhao, "Music artist classification with convolutional recurrent neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[7] M. Volkovs, H. Rai, Z. Cheng, G. Wu, Y. Lu, and S. Sanner, "Two-stage model for automatic playlist continuation at scale," in *Proceedings of the ACM Recommender Systems Challenge 2018*, pp. 1–6, 2018.

[8] J. Dai, W. Liu, C. Ni, L. Dong, and H. Yang, ""multilingual" deep neural network for music genre classification," in *Sixteenth annual conference of the international speech communication association*, 2015.

[9] N. Karunakaran and A. Arya, "A scalable hybrid classifier for music genre classification using machine learning concepts and spark," in *2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*, pp. 128–135, IEEE, 2018.

[10] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2021.

[11] B. G. Patra, D. Das, and S. Bandyopadhyay, "Multimodal mood classification of hindi and western songs," *Journal of Intelligent Information Systems*, vol. 51, no. 3, pp. 579–596, 2018.

[12] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[13] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music

mood regression with deep recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5412–5416, IEEE, 2014.

[14] T. Lidy, A. Schindler, *et al.*, "Parallel convolutional neural networks for music genre and mood classification," *MIREX2016*, 2016.

[15] L. Trottier, P. Giguère, and B. Chaib-draa, "Parametric exponential linear unit for deep convolutional neural networks," 2018.