Semantic Feature Extraction using Feed-Forward Neural Network for Music Genre Classification

Danyal Imran , Hina Wadiwala, Muhammad Atif Tahir, Muhammad Rafi

***Abstract*—Music genre is a conventional category that identifies some piece of music as belonging to a shared tradition or set of conventions characterized by similarities in form, style or subject matter. Traditional method of genre classification tends to extract features and use them to predict labels. These features are independent of each other and do not provide meaning to music genre classification process. In order to achieve semantic meaning of features, feed-forward neural network model with stochastic gradients descent and back propagation algorithm with the categorical cross entropy loss function is investigated in this paper. The main objective is to identify complex patterns that can help in music genre classification. Experiments are performed on AMG1608 dataset and results have indicated significant performance gains when compared with existing approaches.**

I. Introduction

Music is a language that speaks to everyone is their own way. It varies from tradition to culture and is filling the online databases rapidly. It is likely that every music recorded in history will be available online [1]. As these databases are growing enormously, much effort is required to retrieve them accurately. Due to such reasons, Music Information Retrieval (MIR) systems are heavily in demand.

Research in MIR started since 2002 and from then on, many systems have been evaluated with different strategies, yet accuracy is still a measure that is hard to achieve with live systems. Current systems have not adopted the automated strategy and are still under a manual phase, where the user searches for the music through a title, lyrics, singer name, etc. Since the repository is manually marked, there are many margins for errors that the system can not handle and yield a negative result for the user.

Development in MIR systems have targeted objectives such as classification, clustering, tagging, ranking or annotation using a mainstream strategy i.e. to extract low-level descriptors, generate mid-level, high-level or hand crafted descriptors which are summarized as feature vectors to achieve their task.

Music is a mixture of art, concepts, traditions, instruments and melody, therefore poses a challenge while automatically trying to classify the genre of the music [2]. Therefore, low-level and generated descriptors are not enough to classify music accurately since they have reached a standard benchmark. The transition in MIR systems came with the advancements in Deep architecture models that overwhelmed the naive three-step strategy [3]. The advantage that deep architecture offered was that we could generate and learn semantic features by using a combination of many descriptors to make better predictions. Recent studies on deep architecture [4], [5], [6], [7] have shown advantages over other MIR systems.

In this paper, we present a feed-forward neural network to learn semantic features to predict music genre. The aim of the model is to analyze semantic features to improve music genre classification whilst minimizing performance loss over other datasets using stochastic gradients descent and back propagation algorithm. Music data is passed to the extractor that extracts low- level descriptors which are then dimensionally reduced using univariate feature selection algorithm to identify the degree of linear dependency between two random variables and capture any kind of statistical dependency using F-test. Input music once transformed into semantic features through the network, are then mapped to the output layer of the model, which dictates the probabilities of each genre, the output neuron with the maximum probability determines the genre of the music at hand. The idea of back propagation algorithm and stochastic gradients descent algorithm is to reduce error/loss of the classification result by taking advantage of the model architecture by moving similar genre closer whilst moving dissimilar genre farther away from one another. Experiments are performed on AMG1608 dataset, which consists of 1000 songs, each belonging to an individual genre from 10 distinct genres.

The rest of the paper is organized as follows. Section II reviews the related work. Section III discusses presented model of Feed-Forward neural network to learn semantic features followed by experiments and results in Section IV. Section V concludes the paper.

II. Literature Review

Music Genre classification dates back to 2002, when G. Tzanetakis and P. Cook open-sourced their work and MARSYAS framework online [1]. The MARSYAS framework is a low-level feature extractor for audio files that aided researches in working with actual music files. Since then many researchers have worked with different approaches to resolve this problem yet there seems to be no actual system that implements search by music genre, which is highly in demand by music enthusiast. After MARSYAS framework was open sourced, Shen et al. [8] used a combination of acoustic features with the aid of MARSYAS framework to generate a 25-dimensional reduced feature vector. Later being used with neural networks to perform a nonlinear dimensionality reduction. The output generated a single node in the last layer of the network that predicted the genre that the music corresponded to. However, there were several problems faced while performing this task such as Accuracy, Taxonomy etc. [9]. Approaches focused on extracting acoustic features, generating hand-crafted features,

or a dimensional model to map human psychology and physiology to portray music genre in a particular music clip.

Classifying music data is a well known problem and has been massively researched throughout different parts of the world. Techniques from signal processing such as handcrafted features (Cepstrum, Fast Fourier Transform) and modern learning algorithms like Support Vector Machines (SVM), Decision Trees and Deep Neural Networks such as Convolutional Neural Network (CNN), Restricted Bolzmann Machines (RBM) have been applied to solve this problem [20]. Furthermore, variant techniques have also been applied to reduce model's variance using AdaBoost, Bagging, Voting and Aggregation [1], [32] for the same purposes.

Timbre and beat related characteristics acoustic information was discovered by Paradzinets at el. [14]. The features were extracted using Piecewise Gaussian Modelling (PGM) of human auditory filter. To accomplish this, the PGM features we applied to a critical band filter that had equal loudness and sensation. The wavelet transforms were used to get a 2-dimensional beat histogram to extract beat related features. The nodes that possessed relative amplitude of their harmonics were used to fetch timbre related features.

Musical pieces were divided into segments of frames, which were by the extractor to extract MFCC for each segment. The segments were parsed into Gaussian Mixture model (GMM) to predict the genre [19].

The idea to classify music came from the field of digital signal processing since music in its raw form are signals. Therefore, features such as Spectral Centroid (SC), Spectral Flux (SF), Mel-frequency Cepstral Coefficient (MFCC), Zero Crossing Rate (ZCR), etc. were adapted into frames. Although, MIR systems recommend the usage of low-level descriptors [10], [11], [12], there are other systems that recommend hand-crafted features, as they depict more meaning but are harder to optimize. Neural Networks models were trained using Rectified Linear Unit (ReLU) activation function with Stochastic Gradients Descent algorithm to minimize loss with a feedback backpropagation algorithm to tune the weights of the gradient [13]. Hidden layers were trained to with a bias to handle dying neuron problem. The implementation was tested and validation on ISMIR 2004 and GTZAN dataset.

Multiple coefficients were extracted from MFCC to form biological descriptors such as Gamma tone Cepstral Coefficients (GFCC) by Valera and Alias [15]. The results proved that GFCC provided better results for instrumental music over MFCC. Furthermore, GTZAN dataset was tested and validated by Feng using the five layer Restricted Boltzmann Machine (RBM) [16]. The results were completely off but provided a complete new paradigm for researchers to work with. Li et al. [17] followed the above traits by implementing a Convolutional Neural Network (CNN) that was trained using majority voting ensemble rule of a numerous Decision Tree (DT) classifiers. The result of his work was reported at 84% [18]. Li et al. further improved by implementing a Convolutional Deep Belief Network (CDBN) whose objective was to solve artist and genre classification [21]. Dieleman used the above model using pre-trained and engineering features with a vast dataset aimed at solving artist recognition and music genre classification task [22]. Russell and Thayers [23]. model has a great impact as it allowed music to be mapped onto a two dimensional coordinate system. To gain better results, features had to be mapped according to a human brain, this Cognitive analysis of music data was what researchers started to target. Acoustic features would be combined to generate features according to the activation function and the back propagation algorithm to tune stochastic gradients descent algorithm to tune the model to reduce classification errors. Furthermore, some researchers opted a text based approach by combining bag of words with Gaussian super vector (GSV-UBM), to estimate music genre using F0 estimation of music signals. Researchers tried to link genre with other musical metrics and tried enumerations of techniques such as emotion recognition, artist profiling, instrument detection and much more since they believed that a music genre is composed of similar features with respect to other musical domains. Thisismyjam was a social website that allowed user to post their favorite song to learn user traits regarding music [24]. The researches soon made their work open source after they had completed their analysis.

III. Feed-forward neural network model to learn semantic features

In this section, we will discuss feed forward neural network model to learn semantic features. The main objective is to improve music genre classification. We will first discuss preprocessing steps followed by description of feed-forward neural network model. Overview of model is shown in Figure .

Music files in AMG1608 dataset are in .au format. Data in any MIDI format does not provide much insight and cannot be input to any machine learning algorithms and must therefore be passed to an extractor for them to make sense. All features are 2 dimension vectors with varying lengths but are normalized to equal lengths by allowing classification of the first 2 minutes of each piece of music.

A. Feature Scaling

Music features are range from very large to very small number. Difference in such features lead to improper classification task at hand. Therefore, standardizing the range of independent features of data is achieved by normalizing data from range of 0 to 1. This leads to emphasis of feature correlation.

B. Feature Extraction

1) Mel-frequency Cepstral Coefficients (MFCC): These are coefficients that make up a Mel-frequency Cepstral (MFC: representation of the short-term power spectrum of sound).

2) Spectrogram: A spectrogram is a series of short term Discrete Fourier Transform (DFTs), which displays the magnitudes of the audio signal from 0 Hz to its peak frequency.

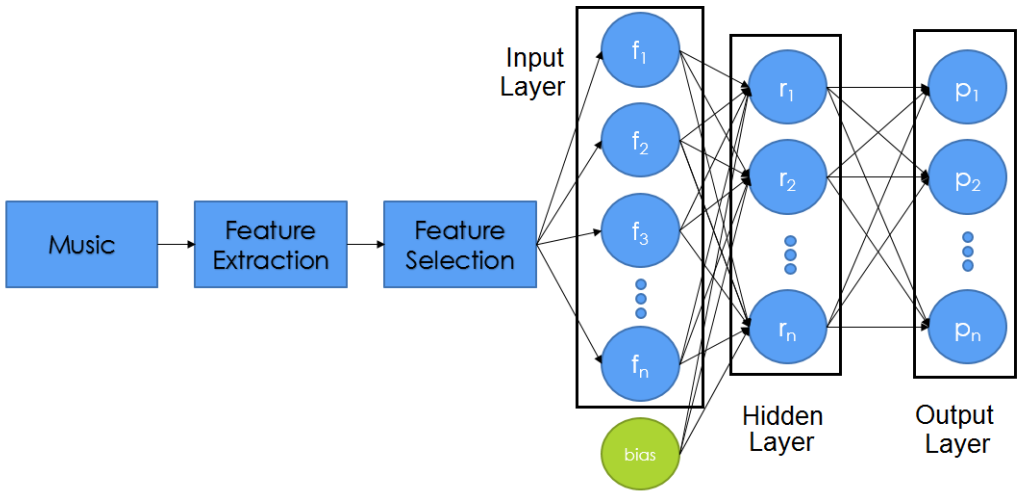

Fig. 1. Feed Forward Neural Network for Music Genre Classification.

3) Chromagram: Chromagram are closely related features to the twelve different pitch classes. These are powerful tools for analyzing music whose pitch can be meaningfully categorized. The twelve pitches are as: C, C#, D, D#, E, F, F#, G, G#, A, A#, B.

4) Spectral Roll-off: Spectral Roll-off is the steepness of a transmission function with frequency. It is helpful in determining the constant gradient at frequencies well away from the cut-off point of the frequency curve. Spectral Roll-off is defined mathematically as

$$\sum_{n=1}^{R_t} M_t^n = 0.5 \times \sum_{n=1}^{N_t} M_t^n \tag{1}$$

where M_t^n represents the magnitude of a particular frame of a signal.

5) Spectral Bandwidth: Spectral Bandwidth is the difference between the upper and lower frequencies in a continuous set of frequencies.

$$SB = \frac{V_o}{V_i} = \frac{1}{1 + iwRC} \tag{2}$$

where R, C represents Resistance and Capacitance respectively. i and w are weighting parameters adjusted by the signals. V_o and V_i represents output and input voltage respectively.

6) Zero Crossing Rate (ZCR): Zero Crossing Rate is the rate of sign changes along a single, i.e. the rate at which the signal changes from positive to negative or vice versa. It is one of the key feature in music information retrieval and speech recognition domain. ZCR is formally defined as follows

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}_{\mathbb{R}_{<0}}(s_t s_{t-1}) \tag{3}$$

where s is a signal of length T and $\mathbb{I}_{\mathbb{R}_{<0}}$ is an indicator function.

7) Harmonic: Harmonic analysis is concerned with the representation of functions of signals as the superposition of basic waves, and the study of and generalization of the notions of Fourier series and Fourier Transform.

8) Percussive: Percussive analysis aims to separate the non-pitched instruments and apply Discrete Fourier Transform to generate a signal irrespective of pitched instruments.

9) Tonnetz: Tonnetz is a conceptual lattice representing tonal space. Tonnetz maps the 12 pitch classes on the tonal space to generate a lattice graph.

10) Spectral Centroid: Spectral Centroid (SC) is used to characterize a spectrum. It indicates the center of mass of the spectrum. It has connection with the "brightness" of a sound. Let $x(n)$ be weighted frequency value from a total of N bins and $f(n)$ is the center frequency for a particular bin, that Spectral Centroid is defined as

$$SC = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{4}$$

11) Root Mean Square Energy (RMSE): Root Mean Square Energy is the measure of loudness of an audio signal. RMSE is defined as

$$RMSE = \sqrt{mean(A^2)} \tag{5}$$

where A is the sum of amplitude of the signal over frequency-time domain.

C. Feature Selection

Feature selection is process of selecting a subset of relevant features for use in model construction. Recursive feature elimination method was adopted i.e. the model was trained on initial number of features and weights were acquired for each feature and the model. If the weight of the model was lower than the desired threshold, the feature with the lowest weight was removed. The algorithm was repeated until the model's weight was above the weighted threshold. After the

TABLE I
RECURSIVE FEATURE ELIMINATION RESULTS FOR EACH ITERATION

Number of features	Model's Weight
11	0.639
10	0.571
9	0.613
8	0.704
7	0.698
6	0.767

algorithm was executed, a total of 6 features were left. The algorithm was chosen to reduce models biasness and variance, so that it acts as a stable model with the unseen data. The model is also based on F-test to estimate the degree of linear independency between two random variables and capture any kind of statistical dependency.

The weighted threshold for the model was set to be 0.75, which is the recommended weight for solving any data science problem. Table 2 shows the result of each iteration for the recursive feature elimination algorithm.

D. Classification using Feed-Forward Neural Network

Classification on selected features were performed by the Feed-Forward neural network for 20 epochs. The reason for choosing this classifier is reported in [13]. The classifier is divided into three layers: input layer, hidden layer and the output layer. The input layer holds values for each feature in an individual neuron, since there are 6 features, the input layer has 6 neurons. The hidden layer is the middle layer where which calculates the interactive features from the input layer, ReLU (Rectified Linear Unit) activation function (ReLU) is used (See Equation 6) to remove negative values from neurons for effective gradient propagation. Weight of edges between neurons are changed in each iteration to minimize loss using stochastic gradients descent and back propagation algorithm. Lastly, the output layers consists of 10 neurons since we want to predict the probabilities of each genre using the sigmoid activation function. Dataset for training and testing the classifier is divided as half (500 songs per training and testing) with 10-folds cross validation method. This strategy introduces random distribution which provides an unbaised result due to the fact that no genre is treated as having a heavier weight than any other genre. The training is completed in approximately 6 hours to provide us with a baseline accuracy. Once an acceptable baseline is reached, the classifier is then tested using the other half testing data. This provides the actual real-time accuracy of the classifier.

$$ReLU = bias + \sum_i w_i * v_i \tag{6}$$

where v_i is the input vector and w_i are the weights of neuron. The learning curve and miss classification rate of the model for 20 epochs is showing in Figure 2.

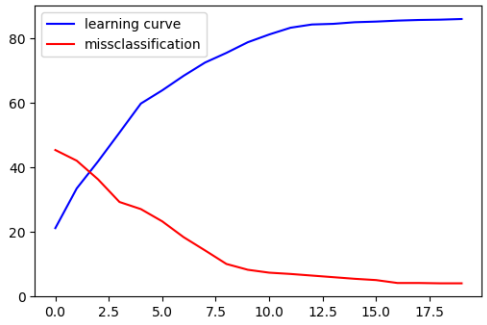

Fig. 2. Feed Forward Neural Network's learning curve.

TABLE II
MUSIC GENRE TYPE AND NUMBER OF SAMPLES.

Genre	Number of Samples
Blues	89
Classical	107
Country	156
Disco	70
Hip Hop	65
Jazz	85
Metal	118
Pop	136
Reggae	121
Rock	53

IV. EXPERIMENTS AND RESULTS

A. Dataset

Dataset used for experimentation is the AMG1608 music dataset. The dataset has 1000 songs, where each song has a single class label from a distinct of 10 genre. The class is a bit imbalanced, especially for some genre, results are reported in Table II.

B. Results and Discussion

After training the neural network on 50% of training data and achieving a baseline accuracy of 71.2%, we tested our results on the other 50% of the data (testing data). The result was quiet good and our presented model able to achieve 77.6% accuracy. The confusion matrix displays the output result for mean result of cross validation from training is shown in Figure 3. It is observed from Figure 3 that Disco, Rock and Blue classes are hard to classify while very high accuracies are obtained for metal and classical categories.

Table III compares the performance of feed-forward NN model with existing approaches. Results have indicated quite significant improvement in performance with compared with methods based on various machine learning and neural network based classifier. Micheal et al [33] achieves very similar performance using neural network but with only 5 targeted classes. Future work aims to investigate ensemble of various neural network classifiers based on Deep Belief Network, Recurrent Neural Network, Radial Basis Function, Convolutional Neural Network, Restricted Boltzmann Machine to improve the overall performance.

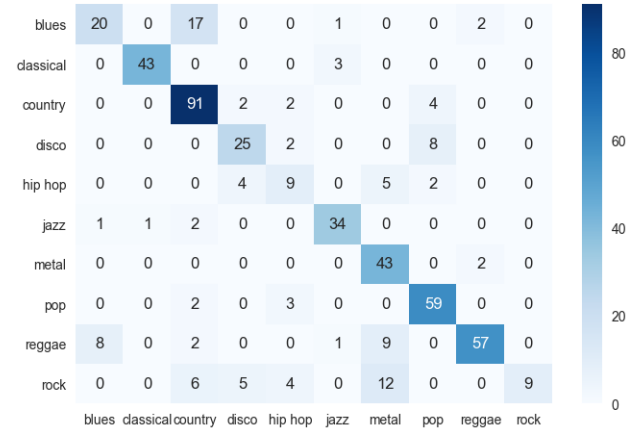

Fig. 3. Feed Forward Neural Network Confusion Matrix Result.

TABLE III
COMPARISON OF FEED FORWARD NEURAL NETWORK MODEL WITH EXISTING METHODS.

	Accuracy
Our Model on 10 target classes	78%
5-NN with Euclidean Distance [1]	61%
KMeans with EchoNest Features [28]	70%
Nave Bayes using m-estimate [21]	73%
Neural Networks on 5 target classes [33]	78%
Deep Convolutional Neural Network [34]	74%
Hierarchical Attention Network on 20 target classes [35]	49%
Majority Voting Classifier on 20 target classes [35]	27%

V. CONCLUSION

In this paper, we have trained a feed-forward neural network on total of 6 features which were selected out of a total of 11 extracted features using univariate feature selection algorithm. The simple nature of the model takes advantage of stochastic gradients descent and back propagation algorithm to move similar genres closer and dissimilar genres far apart whilst minimizing loss using the categorical cross entropy loss function. AMG1608 dataset was used to experiment the model which provided a accuracy of 78% which is comparable with existing approaches.

REFERENCES

[1] G. Tzanetakis and P. Cook. “Musical genre classification of audio signals”. IEEE Transaction on Speech and Audio Processing, 10(5):293-302, 2002.
[2] T. Li, M. Ogihara and Q. Li. “A comparative study on content-based music genre classification”. SIGIR, pages 282-289, 2003.
[3] Y. Bengio, “Learning deep architecture for AI”. Foundations and Trends in Machine learning, 2:1-127, 2009.
[4] P. Hamel, S. Lemicuz, Y. Bengio and D. Eck. ”Temporal pooling and multiscale learning for automatic annotation and ranking of music audio”. International Society of Music Informaiton Retrieval, 2010.
[5] P. Hamel, S. Lemicux, Y. Bengio and D. Eck. “Temporal pooling and multiscale learning for automatic annotation and relevancy of music audio”. International Society of Music Information Retrieval, 2011.
[6] S. Dielemann P. Brakel and B. Schrauwen. ”Audio-based music classification with a pretrained convolutional network”. International Society of Music Information Retrieval, 2011.
[7] J. Nam, J. Herrera, M. Slaney and J. Smith. “Learning sparse feature representation for music annotation and retrieval”. International Society of Music Information Retrieval, 2012.
[8] J. Shen, J. Shepard and A.H.H Ngu. “Towards effective contentbased music retrieval with multiple acoustic feature combination”. IEEE Transaction on Multimedia, 2006.
[9] A. Livshin, G. Peeters and X. Rodet. “Studies and Improvements in Automatic Classification of Musical Sound Samples”. Ircam - Center Pompidou, 2014.
[10] G. Li and A.A. Khokhar. “Content-based indexing and retrieval of audio data using wavelets”. International Conference on Multimedia and Expo, 2010.
[11] U. Nam and J. Berger. “Addressing the same but different but similar problem in automatic music classification”. International Symposium on Music Information Retrieval, 2011.
[12] X. Valero and F. Alias. “Gammatone Cepstral coefficients: biologically inspired feature for non-speech audio classification”. IEEE Transaction on Multimedia, 2012.
[13] S. Sigtia and S. Dixon. “Improved music feature learning with deep neural networks”. IEEE International Conference on Acoustic, Speech and Signal Processing, 2014.
[14] Paradzinets. A, Kotropoulos. C and Chen. L. “Multiexpert system for music genre classification”. Research report, 2009.
[15] S. Golder and B.A. Huberman. “The structure of collaborative tagging systems”. 2006.
[16] T. Feng. “Deep learning for music genre classification”. University of Illinois, 2014.
[17] T. Li, A.B. Chan and A. Chun. “Automatic musical pattern feature extraction using convolutional neural network”. International Conference Data Mining and Application, 2014.
[18] P. Hamel and D. Eck. “Learning features from music audio with deep belief networks”. International Society of Music Information Retrieval, 2010.
[19] Ezzaidi. H and Rouat. J. “Automatic music genre classification using divergence and average information measure”. Research report of the world academy of science, engineering and technology, 2006.
[20] S. Sigtia and S. Dixon. “Improved music feature learning with deep neural networks”. Acoustics, Speech and Signal Processing, IEEE International Conference on, pages: 6959-6963. IEEE, 2014.
[21] A. Locoste and D. Eck. “A supervised classification algorithm for note onset detection”. EURASIP Journal on Applied Signal Processing, 2013.
[22] H. Lee, P. Pham, Y. Largman and A.Y. Ng. “Unsupervised feature learning for audio classification using convolution deep belief neural networks”. Neural Information Processing Systems, 2009.
[23] J.A. Russell. “A circumflex model of affect”. Personality and Social Psychology, 39(6):1161-1178, 2015.
[24] A. Jansson, C. Raffel and T. Weyde. “This is my jam - data dump”. International Society of Music Information Retrieval, 2015.
[25] D. Diekreoger. “Can music lyrics predict genre”. Stanford University, 2012.
[26] Z. Fu, G. Lu, F.M. Ting and D. Zhang. “A surver of audio based music classification and annotation”. IEEE Transaction on Multimedia, 2011.
[27] M.I. Mandel, D. Eck and Y. Bengio. “Learning tags that vary with songs”. International Society for Music Information Retrieval, 2010.
[28] J. C. Wang, H. S. Lee, H. M. Wang and S. K. Jeng. “Learning the similarity of audio music in bag-of-frames representation from tagged music data”. International Society of Music Information Retrieval, 2011.
[29] L. Su, C. C. M. Yeh, J. Y. Liu, J. C. Wang and Y. H. Yang. “A systematic evaluation of the bag-of-frames representation for music information retrieval”. IEEE Transactions on Multimedia, 2014.
[30] E. J. Humphrey, J. P. Bello and Y. LeCun. “Moving beyond feature design: deep architectures and automatic feature learning in music informatics”. International Society for Music Information Retrieval, 2012.
[31] Boyang Gao. “Contributions to music semantic analysis and its acceleration techniques”. Ecole Centrale de Lyon, 2014.
[32] J. Bergstra, N. Casagrande, D. Erhan, D. Eck and B. Kegl. “Aggregate feature and adaboost for music classification”. Machine Learning, 65(2-3): 473-484, 2006.
[33] Micheal Haggblade, Yang Hong, and Kenny Kao. “Music Genre Classification”. International Society of Music Information Retrieval, 2006.
[34] Keunwoo Choi, Gyorgy Fazekas, and Mark Sandler. “Transferring Learning for Music Classification and Regression Tasks. Center for Digital Music”, EECS, 2017.
[35] Alexandros Tsaptsinos. “Music Genre Classification by Lyrics using a Hierarchal Attention Network”. Stanford University, 2017.

AUTHORS

Danyal Imran
School of Computer Science
National University of Computer and Emerging Sciences
Karachi Campus, Pakistan

Hina Wadiwala,nayyar.hussain@faculty.muet.edu.pk
School of Computer Science
National University of Computer and Emerging Sciences
Karachi Campus, Pakistan

Muhammad Atif Tahir,atif.tahir@nu.edu.pk
School of Computer Science
National University of Computer and Emerging Sciences
Karachi Campus, Pakistan

Muhammad Rafi
School of Computer Science
National University of Computer and Emerging Sciences
Karachi Campus, Pakistan