Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 8, July, 2021:7293 – 7305

Research Article

Music Genre Classification Using Deep Learning Techniques

Dr. G.JawaherlalNehru[1], Dr.N.Satheesh[2], Dr.T.Poongothai[3], Dr. B.Rajalingam[4],
Dr. R.Santhoshkumar[5], S. Bavankumar[6], Vishnuvardhan Reddy[7]

Abstract

Music Genre classification (MGC) is very important in today's world due to rapid growth in music tracks, both online and offline. In order to have better access to these we need to index them accordingly. Automatic music genre classification is important to obtain music from a large collection. Most of the current music genre classification techniques uses machine learning techniques. In this paper, we present a music dataset which includes four different genres. A Deep Learning approach is used in order to train and classify the system. Here H2O.Deep Neural Network (DNN) is used for training and classification. Feature Extraction is the most crucial task for audio analysis. Mel Frequency Cepstral Coefficient (MFCC) is used as a feature vector for sound sample. The proposed system classifies music into various genres by extracting the feature vector. Our results show that the accuracy level of our system is around 97.8% and it will greatly improve and facilitate automatic classification of music genres.

Keywords: Music Genre classification, Mel Frequency Cepstral Coefficient, Deep Neural Network.

I. INTRODUCTION

The music dataset downloading from online music collections has become a part of the daily life of probably a large number of people in the world. The users often formulate their preferences in terms of genre, such as hip hop or pop or disco. However, most of the tracks now available are not automatically classified to a genre. Given a huge size of existing collections, automatic genre classification is important for organization, search, retrieval, and recommendation of music.

Throughout computer science, the implementation of the Digital Signal Processing (DSP) and Artificial Intelligence (AI) principles has now become very significant. Musical classification uses the artificial intelligence algorithm to categorize a musical file according to its type. It informs

[1],[4,5,6] Associate Professor, [3]Professor & Head, [4]Assistant Professor
[1,2,5,6] Department of CSE, St. Martin's Engineering College, Secunderabad, Telangana
[2]Department of IT, Annamalai University, Tamilnadu
[3] Swarna Bharathi Institute of Science & Technology
[1]gjnehruceg33@gmail.com

us whether classical pop music is or not. This is very helpful for people who want to arrange the songs on their storage devices, etc

The Music classification have been considered as a very challenging task due to selection and extraction of appropriate audio features. While unlabeled data have been readily available music tracks with appropriate genre tags is very less. Music genre classification is composed of two basic steps: feature extraction and classification. In the first stage, various features are extracted from the waveform. In the second stage, a classifier is built using the features extracted from the training data. There has been many approaches that are used for the classification of music into different genre. With huge amount of music available in the internet it is needed for an automatic music genre classification. Each implementation uses various types of feature extraction. Some can take the timber, rhythm etc. as the classifying parameter, while some others take pitch, timber, beat etc. The types of features extracted varies from person to person. The Deep Neural Network (DNN) is a most widely used in classification problems and it is helpful in training huge database.

We propose a novel approach for the automatic music genre classification using DNN. The features from the music are extracted. They are called as Mel Frequency Cepstral Coefficients (MFCC) for each song. They are obtained by taking the Fourier transforms of the signal, then taking the logarithmic of the power values and then taking the cosine transforms. The detailed explanation will be done in the forthcoming sessions. These extracted features then acts as the inputs to the neuros for training. For our work, we analyze music from four various genres. The whole implementation is done in R programming language. The average accuracy obtained by using the MFCC feature vectors is 97.8%.

The objective of this work is to classify music genres by analyzing their characteristics. This chapter is organized as follows: Section III describes the music genre classification system. Section IV presents feature extraction procedure. Section V explains the proposed method for MGC. Experiments and results are provided in Section VI. Finally, the chapter is summarized in Section VII

II. LITERATURE REVIEW

Gursimran and Neha [1] proposed an automatic music classification system using SVM and Back Propagation Neural Network (BPNN). MFCC are calculated as features to characterize audio content. SVM and BPNN learning algorithms have been used for the classification of genre classes of music by learning from training data. Experimental results show that the accuracy of classification of music genre for SVM is 83 % and for BPNN is 95%. The performance of BPNN is better than the SVM model.

Lima et al. [2] assessed the music genre classification using spectrograms taken from the original signal, percussive content signal, and harmonic content signal. The rationale behind this is that classifiers obtained from these three different representations of the signal may be complementary to each other. LBP texture features are used to represent the spectrogram content, and the classification is done by SVM. The spectrogram images were zoned, by taking into account a perceptual scale, and a specific classifier was created for each zone. Finally, classifiers outputs were combined to get the final decision. This approach shows 88.56% recognition rate.

Martins et al. [3] proposed a new music dataset called BMD (Brazilian Music Dataset) containing 120 songs labeled in 7 musical genres, FoFF, Rock, Repente, MPB, Brega, Sertanejo and Disco. The authors evaluated proposed features on both datasets: GTZAN and BMD. The proposed approach achieved average accuracy (after 30 runs of 5-fold-cross-validations) of 79.7% for GTZAN and 86.11% for the BMD.

Panwar et al. [4] proposed CRNN model by combining CNN and Recurrent Neural Networks (RNN). CNN extracts local features at different levels of hierarchy using kernels, and RNN discover the global features to understand the temporal context. CRNN architecture utilizes the benefits of both CNN and RNN structures. The authors used CRNN structure on MagnaTagA Tune dataset. The AUC-ROC index for the proposed architecture is 0.893 which shows its superiority rather than traditional structures on the same database.

Muhammad and Zain [5] proposed the machine learning algorithms that predict the genre of songs using k-NN and SVM. The authors presented comparative analysis between k-NN and SVM with dimensionality reduction and then without dimensionality reduction via Principal Component Analysis (PCA). The MFCC is used to extract information for the data set. In addition, the MFCC features are used for individual tracks. From the results it is found that without dimensionality reduction both k-NN and Support SVM gave more accurate results compared to the results with dimensionality reduction. Overall the SVM is much more effective classifier for classification of music genre. It gave an overall accuracy of 77%.

Thiruvengatanadhan [7] proposed an automatic music genre classification system using SVM. Tempogram is calculated as features to characterize audio content. SVM learning algorithm has been used for the classification of genre classes of music by learning from training data. Two nonlinear support vector machine classifiers are developed to obtain the optimal class boundaries between classic and pop, pop and rock by learning from training data. Experimental results show that the SVM learning method performs well with an accuracy rate of 95%.

Hareesh [9] proposed two approaches to classify music automatically by providing tags to the songs present in the user's library. The first approach uses CNN which is trained end to end using the features of spectrogram images of the audio signal. The second approach uses various Machine Learning (ML) algorithms like Logistic Regression, Random forest etc, where it uses hand-crafted features from time domain and frequency domain of the audio signal. The manually extracted features like MFCC, Chroma Features, Spectral Centroid etc are used to classify the music into its genres using ML algorithms like Logistic Regression, RF, Gradient Boosting (XGB), SVM. By comparing the two approaches they came to a conclusion that VGG-16 CNN model performs well with 89% accuracy.

III. PROPOSED SYSTEM

The classification system of the music genre consists of three stages: pre-processing of signals, extraction of characteristics, and classification. The pre-processing operations such as preemphasis, framing, and windowing are performed on music information to allow extraction of features. The mathematical algorithm analyzes the sound and classifies the music genres using the method of signal processing and machine learning. Figure.1 displays the block diagram for the proposed method

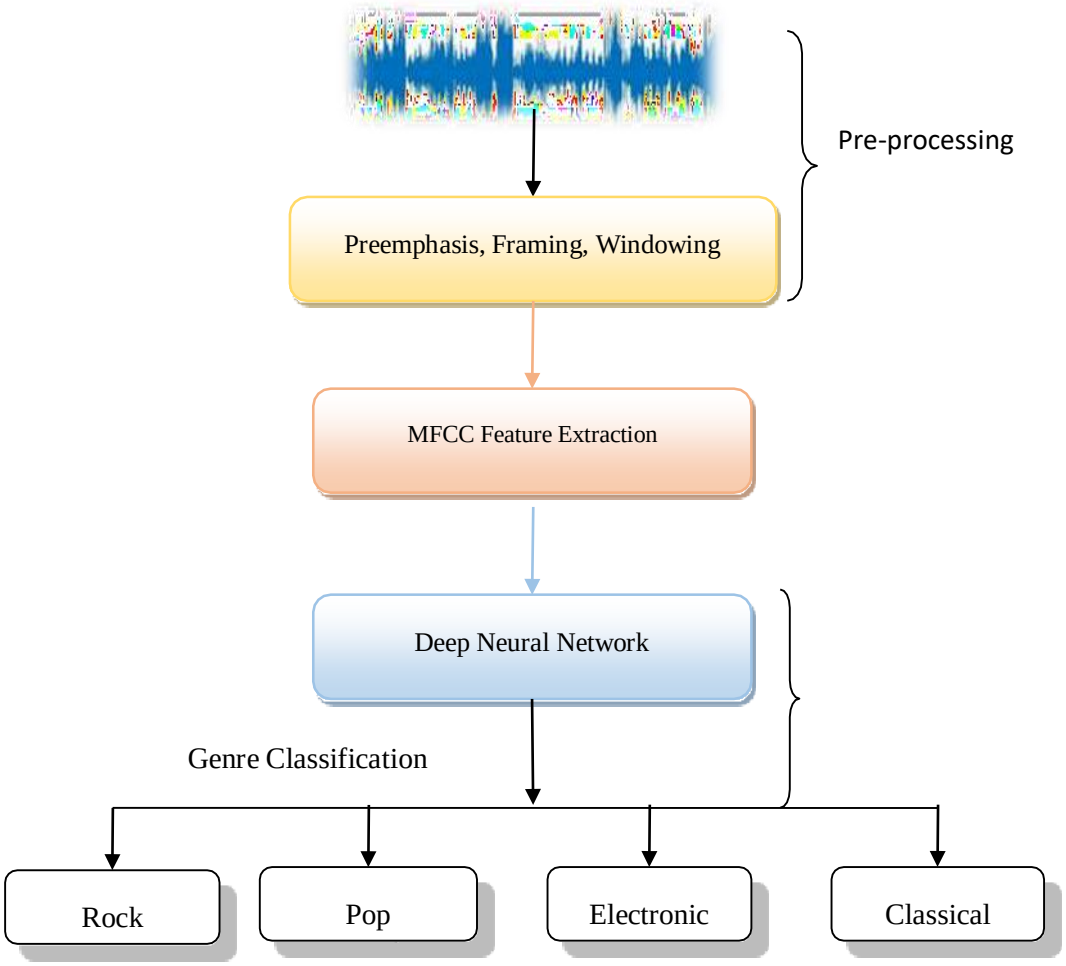


Fig. 3.1 Proposed Music Genre Classification System

FEATURE EXTRACTION

Pre-processing

The pre-processing stage is common for all the feature extraction techniques. The raw digital music signal is pre-processed to enhance the recognition system performance or to prepare the music signal for the extraction stage of the application. Preemphasis filtering, framing, and windowing are the main steps in pre-processing.

Feature Extraction

A sampling rate of 22.5 kHz and 16 bit monophonic PCM format are the preferred properties of the music signals. In this work, MFCC features are used with delta and acceleration

coefficients. Then, the data is segmented into fixed length by 40ms frames with an overlap of 50 percent between adjacent frames using hamming window. For each frame, first 13 Cepstral coefficients values are used. MFCC values are extracted from GTZAN data using R library.

III PROPOSED MUSIC GENRE CLASSIFICATION

This chapter provides an efficient method for the recognition of music genres based on deep learning model. DNNs are considered to be predominant classifiers capable of modeling highly non-linear input-output relationships. After the feature extraction process it is important to classify the input music signal. In the function space, the classifier denotes decision boundaries, separating different sample groups from each other. Multi label feed-forward DNN architecture to various feature representations are generated using signal processing methods.

Deep Neural Network

The architecture of the deep learning system is designed to have non-linear units processing information and different levels of data abstraction are performed in several layers in these models[10]. A multilayer feed forward of the neural network consists of an input layer to match the feature space, followed by several hidden layers, and a classification layer to match the output space.Each neuron is named as a unit in a typical neural network. A layer is considered as a set of neurons in a stack in a neural network. A layer can contain *n* number of nodes.

A typical neural network model has a single input layer and one or two hidden layers can be connected directly from the input layer to the output layer receiving data. The DNN input and output displayed in Figure 2. Each neuron in one layer is connected to the neurons of each other layer. In this work, the number of neurons in the input layer depends on the sizes of the input vector, while the number of the output layer neurons is equivalent to the number of genres to be considered.

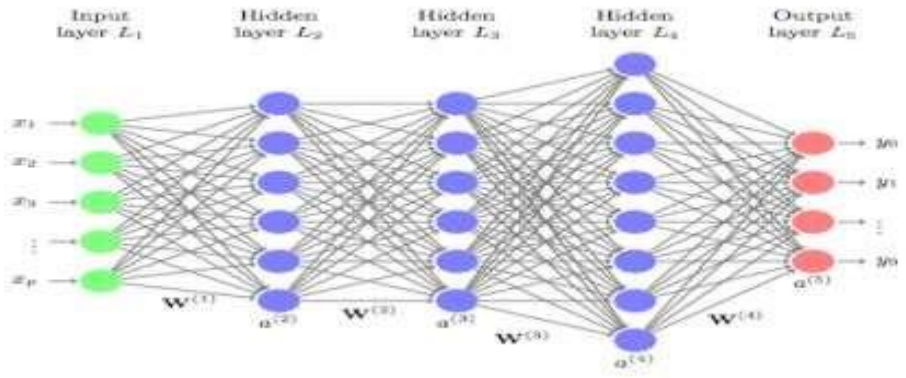
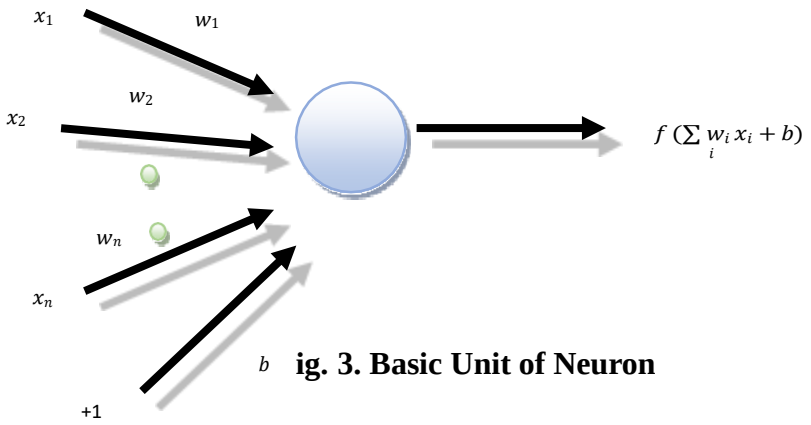


Fig. 2 Deep Neural Network

Within this model, bias value is assigned as 1 within any neural network, which is typically assigned as 1 to prevent invalid network results. In addition, the learning rate is assigned as 0.15

by default and later changed randomly by trial and error to obtain varying model outcomes. The initial weight of the nodes can be assigned randomly and changed by the network during back propagation by calculating error rate and updated periodically after every epochs.

The inputs and outputs of the model's units follow the basic logic of the single neuron. The basic unit in this model is a neuron which is shown in Figure 3.



ig. 3. Basic Unit of Neuron

Feed forward Neural Networks

The simplest type of layer is the feed forward layer. The feed forward layer performs the following transformation:

$$f_i(h_{l-1};\theta_l) = g(W_l h_{l-1} + b_l)$$

where h_{l-1} is the output from the previous layer, W_l is a weight matrix of size $M \times N$ that transforms an M dimensional input h_{l-1} to an N dimensional vector, b_l is an N-dimensional bias vector and $\theta_l = \{W_l , b_l\}$ are the layer parameters. A feed forward layer performs an affine transformation of the input, followed by the application of an element-wise non-linearity. A network comprising only feed forward layers is called a feed forward neural network or a deep neural network (DNN). The non-linearities are essential for learning complex non-linear mappings between the inputs and the outputs.

Activation Functions

Choosing an activation mechanism for a neuron has a variety of choices. Non-linear activation functions provide the ability of neurons to learn non-linear patterns.

The most commonly used activation functions are the sigmoid, tanh and the more recent, Rectified Linear Unit (ReLu).

***Sigmoid*:** Sigmoid activation function is shown in Figure 4 and defined by

$$\sigma(x) = \frac{1}{1+e^{-z}}$$

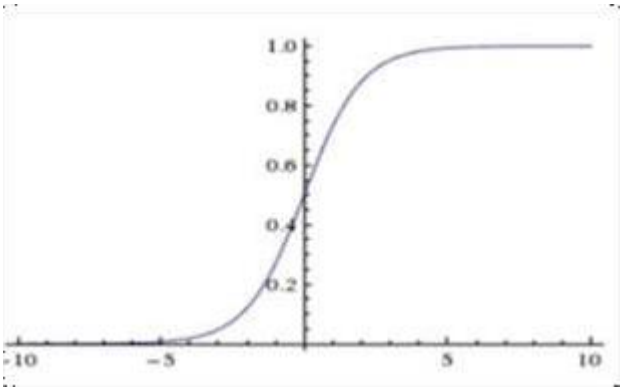


Fig. 4 Sigmoid Function

The sigmoid method takes every real number and positions it within the range of 0 and 1. It means that extremely small numbers are converted to 0 and large numbers are converted to 1. This property can lead to an unintended concept called "saturation" as extremely small or large numbers are always depicted by 0 or 1.

Tanh: It is shown in Figure 5 and given by

$$tanh(x) = 2\sigma(2x) - 1$$

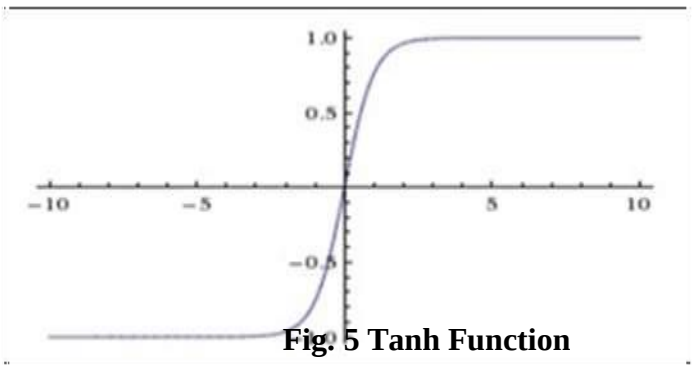


Fig. 5 Tanh Function

The $tanh$ function takes any real valued number and places it into the range of −1 and 1. Similar to the sigmoid, it is also verdict to saturation. However, their output is zero-centered.

ReLu: This function is shown in Figure 6 and defined by

$$f(x) = max(0, x)$$

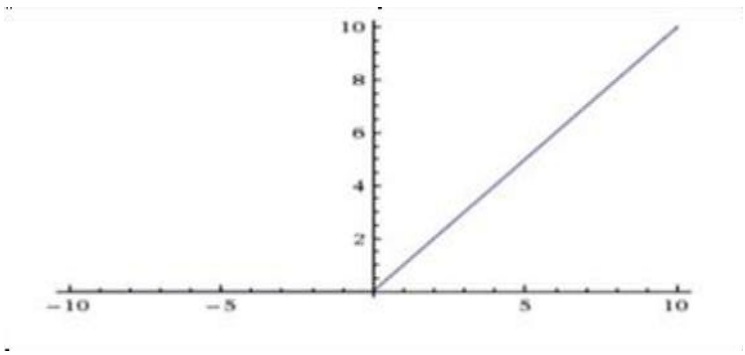


Fig. 6 ReLu Function

The ReLu takes any input and thresholds it at 0. In comparison to the sigmoid and *tanh*, it is relatively computationally inexpensive. They have recently gained popularity due to their fast performance in computer vision.

H2O-Deep Learning

Deep learning (DL) is a system consisting of a collection of methods that identify raw data as meaningful information fed to the machine. Deep convolution networks divided into various processing layers for learning and representing the information [6,8]. It has multiple levels of abstraction for processing images, video, speech, and audio.

H2O's deep learning building design has many features, including supervised training protocol, memory-efficient Java implementation, optimized learning, and related CRAN packages. Clusters of machine learning nodes can be used to train the entire data set, but instead to automatically shuffle training examples for each iteration locally. The framework continues to support softmax methods to prevent overfitting.

R machine learning library H2O Deep Learning is used for implementing DNN in the proposed work. For smarter applications, H2O is fast, scalable, open-source machine learning and profound learning.

Regularization

H2O's Deep Learning uses *l*1 and *l*2 regularization as well as the concept of 'dropouts' to avoid overfitting. *l*1 limits the absolute weight value. It makes a lot of weights 0. *l*2 limits the square weights sum. It makes a lot of weights small. The proportion of input dropout refers to the fraction of features to be removed or omitted from each training record to enhance generalization. For example,' input dropout ratio= 0.1' means that 10% of the input characteristics are lost.

Hidden dropout ratio refers to the fraction of features to be removed or omitted from each hidden layer to generalize. For example, in a DL model with 2 secret layers,' hidden dropout ratios= c (0.1, 0.1)' means that 10% of hidden features are dropped in each hidden layer.

Advanced optimization

H2O features include automatic advanced optimization and manual modes. Manual mode features include annealing of momentum training and learning speeds, while automatic mode features adaptive learning rates.

- **Momentum training:** Momentum computes back propagation by allowing previous iterations to affect the current release. It is possible to avoid local minimum rates and the resulting unrest by using the momentum parameter.

- **Rate annealing:** As the model approaches the minimum chance of oscillation or "optimum skipping," a slower learning speed is required during learning. For H2O, the annealing rate is opposite to the number of samples required to reduce the rate of learning by half.
- **Adaptive learning:** The Adaptive Learning Rate Algorithm ADADELTA dynamically incorporates the advantages of learning rate annealing and momentum learning to prevent slow convergence. Only two parameters ρ and μ are specified, which facilitates the search of the hyperparameter. In certain instances, guided (non-adaptive) learning speeds and adaptive criteria can produce better results but need to test up to 7 parameters for hyper parameters**.**

When the model is based on a topology with many more low or long plateaus locally, it is possible to produce sub-optimal results at a constant learning speed. The first of two adaptive learning hyper parameters is μ. It is like dynamics and concerns the preliminary weight updates in the memory.

The values typically range from 0.9 to 0.999. The second of two hyper parameters ρ is comparable to a learning rate during initial training and momentum during subsequent stages where progress can be achieved.

IV EXPERIMENTS AND RESULTS

Dataset

The dataset used for the analysis is the GTZAN dataset used to test the structures of genre classification. It includes 1000, 30sec album samples, sampling frequency at 16 bit 22050 Hz. Four genres of music such as classical, pop, rock, and electronic are considered in the proposed work. The music server is comprised of 400 metadata audio files. For each genre 100 audio tracks of 30sec long are considered.

The music segments are 16 bit mono in the proposed work. Furthermore, the original rate of sampling is kept without down sampling because the relevant spectral properties have been found in the high frequency range. The data set is divided randomly into three parts for the classification of musical genres: 60% for training purposes, 20% for validation and 20% for testing.

Table.1 DNN Parameters

No. of hidden layers	2
No. of hidden units	350×350
Activation function	ReLu
No. of epochs	20

Proposed Deep Neural Network Parameters

The DNN Parameters used in the proposed MGC system is shown in Table 1. The proposed DNN architecture consists of 2 hidden layers each with 350 hidden units. The traditional activation function used in each hidden layer is replaced by Rectified Linear Units (ReLus). The function of

ReLu has faster computation than sigmoid function and it is applied to all layers except output layer which uses softmax output with categorical cross-entropy loss function.

The proposed model is evaluated using four cross fold validation scheme. For each class, per class accuracy is calculated on frame wise level. These scores are derived by dividing the correct number of frames by the total number of frames of that group. Finally, an average of the cross fold accuracy determines the overall score.

DNN has a problem of overfitting with small data. Dropout technique has been used to avoid this problem. The learning rate is set to 0.0001 and initially 10 numbers of epochs is set into the model. The performance of DNN is evaluated by varying the number of epochs up to 60. The momentum is set to 0.2. The input dropout ratio is initialized as 0.2 for both hidden layers. The classification performance is based on accuracy metrics.

Performance Analysis

The classification performance of the system is measured in terms of accuracy. The proposed model achieves a better accuracy when compared to the existing machine learning models. The overall confusion matrix of the DNN model is shown in Table.2.

Table.2 Confusion Matrix of Proposed MGC System

Genres	Classical	Electronic	Pop	Rock	Accuracy (%)
Classical	2268	140	0	0	**93.23**
Electronic	0	2455	40	0	**98.01**
Pop	0	0	2347	168	**93.07**
Rock	0	0	0	2508	**100.00**

The proposed MGC system is trained using H2O deep learning and evaluated with respect to classification error rate. Loss function is used to evaluate misclassification error rate. Figure 7 shows the error rate obtained during training and evaluation. From the scoring history it is observed that, the classification error rate has been gradually reduced with respect to number of epochs.

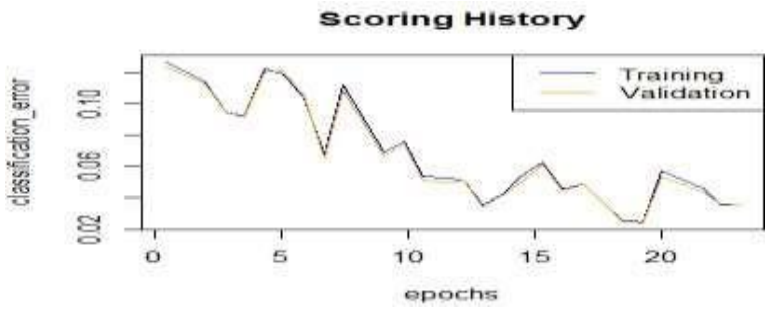


Fig. 7 Classification Error Rate

The frame based accuracy of proposed MGC system using four cross validation is shown in Table.3. H2O deep leaning model with proposed parameters obtained an overall accuracy of 97.8% which is higher than other existing systems.

Table.3 Cross Fold Accuracy

Cross Fold	Frame Based Accuracy (%)
Fold1	**94**
Fold2	**97**
Fold3	**96**
Fold4	**96**

V CONCLUSION

This study aims to classify and recommend songs using acoustic features, extracted by digital signal processing methods and Deep Neural Networks. Study has been conducted over two steps; determining how features that will be used in recommendation are obtained and developing a service that recommends songs to user requests. Firstly, feature extraction has been carried out by means of digital signal processing methods and then H2O.DNN has been trained as an alternative feature extraction. Then acoustic features of songs are used in classification to determine the best classification algorithm and the best recommendation results. This paper presents MGC system using deep neural network model. The performance of H2O deep learning based music classification system has been analyzed with MFCC features and its shows 97.8% of accuracy which is higher than the existing work. The MGC system proves that classification error rate gets gradually reduced.

References

1. Gursimran Kour., and Neha Mehan., (2015), Music Genre Classification using MFCC, SVM and BPNN, International Journal of Computer Applications, vol.112, no. 6.
2. Lima Aguiar., R., Gomes da Costa., Y. M., and Nanni, M., (2016), Music genre recognition using spectrograms with harmonic-percussive sound separation," 35th International Conference of the Chilean Computer Science Society (SCCC), Valparaiso, pp. 1-7.
3. Martins de Sousa, J., Torres Pereira, E., and Ribeiro Veloso, L., (2016), A robust music genre classification approach for global and regional music datasets evaluation, IEEE International Conference on Digital Signal Processing (DSP), Beijing, pp. 109-113.
4. Panwar, S., Das, A., Roopaei, M., and Rad, P., (2017), A deep learning approach for mapping music genres, 12th System of Systems Engineering Conference (SoSE), Waikoloa, HI,, pp. 1-5.
5. Muhammad Asim Ali., and Zain Ahmed Siddiqui., (2017), Automatic Music Genres Classification using Machine Learning, International Journal of Advanced Computer Science and Applications (IJACSA),Vol. 8, No. 8.

6. P.Deepan and L.R. Sudha(2020), "Remote Sensing Image Scene Classification using Dilated Convolutional Neural Networks", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.7, pp.3622-3630, ISSN: 2347-3983.
7. Thiruvengatanadhan, R., (2018), Music Genre Classification using SVM, International Research Journal of Engineering and Technology (IRJET), Vol.05, no.10.
8. P.Deepan and L.R. Sudha, "Comparative Analysis of Remote Sensing Images using Various Convolutional Neural Network", EAI End. Transaction on Cognitive Communications, 2021. ISSN: 2313-4534, doi: 10.4108/eai.11-2-2021.168714.
9. Hareesh Bahuleyan., (2018), Music Genre Classification using Machine Learning Techniques, University of Waterloo.
10. P.Deepan and L.R. Sudha, (2021) "Deep Learning and its Applications related to IoT and Computer Vision", Artificial Intelligence and IoT: Smart Convergence for Eco-friendly Topography, Springer Nature, pp. 223-244, https://doi.org/10.1007/978-981-33-6400-4_11.
11. R. Santhoshkumar, M. Kalaiselvi Geetha, J. Arunnehru, (2017) 'SVM-KNN based Emotion Recognition of Human in Video using HOG feature and KLT Tracking Algorithm, *International Journal of Pure and Applied Mathematics*, vol. 117, No. 15, pp.621-624, ISSN: 1314-3395.
12. R. Santhoshkumar, M. Kalaiselvi Geetha (2019), 'Deep Learning Approach: Emotion Recognition from Human Body Movements', *Journal of Mechanics of Continua and Mathematical Sciences (JMCMS),* Vol.14, No.3, pp.182-195, ISSN: 2454-7190.
13. R. Santhoshkumar, M. Kalaiselvi Geetha (2019), 'Vision based Human Emotion Recognition using HOG-KLT feature*' Advances in Intelligent System and Computing, Lecture Notes in Networks and Systems*, Vol.121, pp.261-272, ISSN: 2194-5357, Springer https://doi.org/10.1007/978-981-15-3369-3_20
14. R. Santhoshkumar, M. Kalaiselvi Geetha (2019), 'Human Emotion Prediction Using Body Expressive Feature', *Microservices in Big Data Analytics, IETE Springer Series*, ISSN 2524-5740, 2019, (Springer), https://doi.org/10.1007/978-981-15-0128-9_13
15. R. Santhoshkumar, M. Kalaiselvi Geetha (2019), 'Emotion Recognition System for Autism Children Using Non-verbal Communication', *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* Vol.8, No.8, pp.159-165, ISSN: 2278-3075.
16. B.Rajalingam, R.Priya, R.Bhavani., Hybrid Multimodal Medical Image Fusion Algorithms for Astrocytoma Disease Analysis. *Emerging Technologies in Computer Engineering: Microservices in Big Data Analytics, Springer*, Vol. 985, 2019, pp. 336–348
17. B.Rajalingam, R.Priya, R.Bhavani., Hybrid Multimodal Medical Image Fusion Using Combination of Transform Techniques for Disease Analysis. *Procedia Computer Science*, Elsevier, Vol. 152, 2019, pp. 150-157
18. B.Rajalingam, R.Priya, R.Bhavani., Multimodal Medical Image Fusion Using Hybrid Fusion Techniques for Neoplastic and Alzheimer's Disease Analysis, *Journal of Computational and Theoretical Nanoscience*, Vol. 16, 2019, pp. 1320–1331
19. B.Rajalingam, R. Santhoshkumar (2020) "Intelligent Multimodal Medical Image Fusion with Deep Guided Filtering", Multimedia Systems, Springer-Verlag GmbH Germany, part of Springer Nature 2020

20. K.P. Sanal Kumar, S Anu H Nair, Deepsubhra Guha Roy, B. Rajalingam, R. Santhosh Kumar " Security and privacy-aware Artificial Intrusion Detection System using Federated Machine Learning" Computers & Electrical Engineering, Volume 96, Part A, December 2021, 107440
21. Dr. B. Rajalingam, Dr. R.Santhoshkumar, Dr. G. Govinda Rajulu, Dr. R. Vasanthselvakumar, Dr. G. JawaherlalNehru, Dr. P. Santosh Kumar Patra "Survey On Automatic Water Controlling System For Garden Using Internet Of Things (Iot)" The George Washington International Law Review, Vol.- 07 Issue -01 April-June 2021.