Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**

CLASSIFICATION OF MUSICAL INSTRUMENTS SOUND USING PRETRAINED ALGORITHMS WITH SUPPORT VECTOR MACHINE

S. Prabavathy[1], R. Selvalakshmi[2]

[1,2]Assistant Professor

[1]Department of Computer Applications, [2]Department of Computer Science

[1]Arulmigu Subramania Swamy Arts and Science College, Vilathikulam, Thoothukudi, Tamilnadu, India.

[2]Kamaraj College (Autonomous), Thoothukudi, Tamilnadu, India.

To Cite this Article

S. Prabavathy, R. Selvalakshmi" **CLASSIFICATION OF MUSICAL INSTRUMENTS SOUND USING PRETRAINED ALGORITHMS WITH SUPPORT VECTOR MACHINE**" *Musik In Bayern, Vol. 90, Issue 10, Oct 2025, pp262-272*

Article Info
Received: 30-09-2025 Revised: 07-10-2025 Accepted: 18-10-2025 Published: 29-10-2025

Abstract

Music can be created through the interaction of various instruments. Music is the universal language that told, the language of the emotions. Humans can identify the music instrument from which it was played, but it is difficult for a machine to do so automatically. In today's decade, classifying music signals from large datasets is a major task; the proposed work classifies music into its respective categories. In this paper, the sounds of musical instruments are automatically classified using spectrogram images created from musical signals. Pre-trained Convolutional Neural Network models, AlexNet and GoogleNet, are used to extract features, and a machine learning model, Support Vector Machine (SVM), is used for classification. This system was tested with 15 different types of musical instrument sounds from three different instrument families: brass, string, and woodwind instruments. In this proposed work, the system achieves 97.6% accuracy by combining GoogleNet with SVM.
Keywords: Musical Instrument Sound Classification (MISC), AlexNet, GoogleNet, Support Vector Machine (SVM).

1. INTRODUCTION

Digital music distribution is now possible through developments in digital storage, audio compression, and enormous network bandwidth expansions. Portable music collections of

thousands of tracks are commonly found on smartphones, computers, and portable music players. Millions of tracks are readily downloadable from digital stores where consumers can buy music. Since the assigned class labels are directly presented to the user and used as a filter, classification is more useful when creating extensive music collections that have been explored. As a result, it is used in an indirect way to recommend music where all labels have similarities based on sophisticated listening habits and to find songs to listen to from the classes where a user is most similar.

Classifier learning and feature extraction are the two main components of classification systems [1]. In the first place, feature extraction is used to report the problem in a way that uses feature vectors or pairwise affinities to represent the samples. Second, classifier learning is primarily used to lower the prediction error and find a mapping from feature space. Unless a different module is used, it is primarily focused on classifying music based on audio signals.

For the purpose of classifying music, this study includes a vast amount of audio features. Additionally, a variety of taxonomies are offered to categorize the audio characteristics. [2] separated the audio characteristics into four subclasses: compositional, semantic, long-term, and short-term features. [3] used benchmark taxonomy to classify audio features used for genre classification into three classes based on pitch, rhythm, and timbre data, respectively. Every taxonomy aims to capture audio characteristics from a particular perspective.

1.1 Musical Instrument Sound Classification

Predicting an instrument's sound, recognizing, retrieving, transcribing, and other methods are all part of Musical Instrument Sound Classification (MISC). Classifying the songs into various genres based on input songs and determining the emotions conveyed by the music are a few examples of MIC. Research on music analysis has been very active. Every kind of instrument has a unique sound and is associated with a particular genre of music. Experienced musical listeners can quickly identify the musical instruments that are present in a music recording, even though the accuracy of identification varies depending on factors like familiarity, the number of instruments played, and the importance of an instrument (inside the music). Humans classify music signals implicitly, but computer systems find this task somewhat difficult. Using digital audio signal processing involves a number of steps, such as producing audio effects, classifying audio, and compressing digital audio signals. In today's world, it is evident how crucial multimedia content management is to audio segmentation and classification. The handling of audio and visual data is where the

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**
main problems occur. It is used for audio classification and has significant applications across many domains [9].

1.2 Musical Instruments Types

Generally speaking, musical instruments are considered to be universal manifestations of human culture. According to archaeology, clay drums and shell trumpets date from the Neolithic Period, while pipes and whistles date from the Paleolithic Period. A longer and earlier deployment of a model is demonstrated by the fact that the ancient city cultures of Mesopotamia, the Mediterranean, India, East Asia, and the Americas are made up of well-equipped assortments of musical instruments.

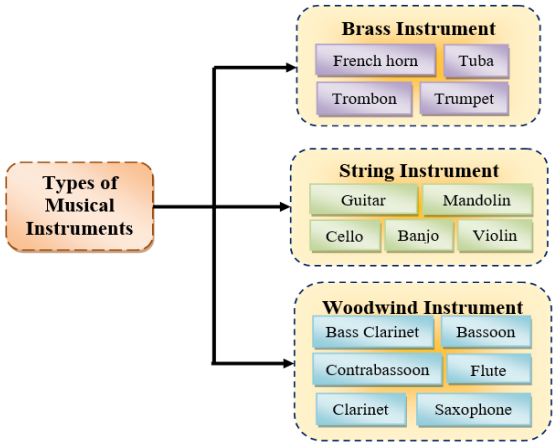

Fig. 1. Musical Instruments Classification (MIC)

Fifteen distinct musical instrument types—from the three families of musical instruments—such as brass, string, and woodwind—are categorized in this work. This proposed work includes the classification of the following musical instrument classes: French horn, Trumpet, Trombone, and Tuba from Brass, Banjo, Mandolin, Cello, Violin, and Guitar from String and Flute, Bass clarinet, Clarinet, Saxophone, Contrabassoon, and Bassoon from Woodwind. The classification of musical instruments is displayed in Fig. 1.

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**

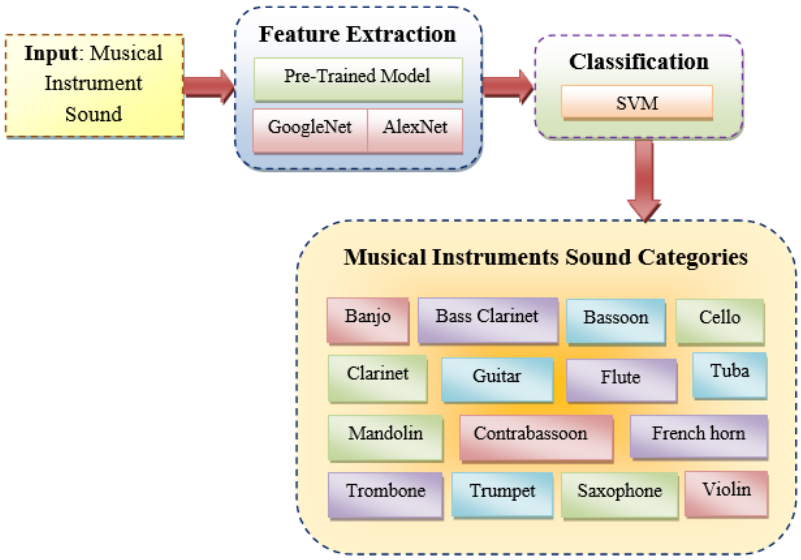

Fig. 2. Block diagram of the proposed work

In this proposed work, the classifier SVM is used to classify the musical instruments, while the pre-trained models, AlexNet and GoogleNet, are used to extract the features from the music signal and generate spectrograms. The block diagram for the suggested work is displayed in Fig. 2.

2. LITERATURE REVIEW

Mid-level representations, in particular, are called time-frequency implications because they compute the amplitudes of a signal in various frequency bands and vary in time and frequency on a large coarser time scale, instead of defining the amplitude of a signal because it has diverse time [4]. Consequently, the predetermined implications sacrifice frequency resolution for temporal resolution. The spectrograms determine the commonly used time-frequency representation. The short-time Fourier transform (STFT) of shorter windows, which frequently overlap, is used to create a signal's spectrogram. Power spectra, which show the power in various frequency bands within a window, may be obtained using the converted windows. A spectrogram is created by combining the successive frames as a matrix.

In [5] the authors described the Indian drum's harmonic behavior and acoustics. The differences between Western drums and the Indian tabla are then described. A random sound cannot be referred to as music since a musical sound has specific wave patterns. Two important factors that make up music are pitch and timber. The fundamental and maximum frequencies of sine waves, known as overtones, make up the notes produced by musical instruments. Higher overtones in a particular sound are intended to be harmonic, but the waveform would not be regular and it would

not be made up of recognizable pitch. A note must therefore have massive overtones that are harmonic in order to sound musical.

In [6] the authors offered a number of models for transcription of drum loops that replicate the variability in intelligent audio clips using a dataset of drum signals. Both SVM and HMM have been used to define the technologies. A higher recognition rate was achieved for some of the more difficult results. However, melody songs are the focus of current speech indexing and music database retrieval efforts. The authors in [7] described an automated drum sound illustration system for real-world musical audio signals. Additionally, it demonstrated how to apply and match templates to identify problems with variations in drum sounds. The results of the experiments suggested that it was possible to accurately identify snare drums and bass in popular music.

The authors in [8] offered a useful tutorial on speaker recognition that outlines the basic steps involved in audio processing and gives a concise explanation of automated speaker analysis. It functions in two modules: identifying a particular speaker or confirming a person's purported resemblance. The basic elements of automated speaker recognition systems and speech computation are shown, along with the developed tradeoffs. A speaker analysis technique that uses LPS frequency characteristics and data-theoretic shape measures to categorize speakers has been proposed.

3. FEATURE EXTRACTION

Sonograms were used to create spectrogram images for the music signal. The resulting spectrograms are used to train the GoogleNet and AlexNet models to perform musical instrument sound classification tasks and extract deep features.

3.1 AlexNet Model

Fig. 6.2 shows a structural design of this network. Three FC and five convolutional (Conv) layers make up its eight learned layers. After that, some of the remarkable characteristics related to network structure are defined. For modeling a neuron's output f as a function of its input x, f(x)= tanh (x) or f(x)=(1+e^(-x))^(-1) are better approaches. The non-saturating nonlinearity f(x)= max (0,x) is faster than the saturating nonlinearities in terms of training time using gradient descent (GD). ReLUs are neurons that exhibit nonlinearity. ReLUs train many instants faster in deep convolutional neural networks (DCNNs) than their corresponding tanh units. The nonlinearity f(x)=| tanh (x)| is said to fit particularly well with their range of contrast normalization and local average pooling.

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**

3.2 GoogleNet

Finding the best way to approximate and hide a convolutional vision network's best local sparse structure with accessible elements is the main goal of an Inception structure. It is clear that convolutional building blocks would be used to develop this model's translation invariance. Finding the best local construction and applying the same procedure spatially is the main idea. An individual should analyze the final layer's association statistics and form groups of units with the highest association in this layer-by-layer development. These gathered clusters are connected to the units of the current layer and further developed as units of subsequent layers. Take into consideration that each unit from the primary layer, which suggests the input image, is gathered inside a filter bank. The appropriate units for lower layers are concentrated in nearby areas. It means that, as advised, a layer of 11 convolutions in the following layer hides the massive clusters that have concentrated in a single area at the end. As a result, it anticipates a smaller number of patches over a large area and a minimum count of spatially distributed clusters that are concealed by convolutions across maximum patches.

Recent iterations of the Inception structure are restricted to filter sizes 11, 33, and 55 in order to eliminate the patch-alignment issues; as a result, the solution relies on necessity and convenience. Additionally, it states that the suggested structure is described as an integration of these layers with matching filter bank results integrated as a distinct output vector creating the input for the next phase. Furthermore, since pooling tasks are primarily necessary for efficient art convolutional networks, this suggests including a backup parallel pooling path for additional advantages. Since the "Inception modules" are stacked on top of each other, the statistics of the resulting association are limited to various objects: since maximum layers have captured features of maximum abstraction, spatial concentration should be decreased by suggesting a ratio of 33 to 55 convolutions, which must be improved as one moves up the layers.

4. CLASSIFIER

The deep features from the pre-trained models AlexNet and GoogleNet are classified using the robust machine learning algorithm known as support vector machine (SVM) as a classifier.

SUPPORT VECTOR MACHINE

One machine learning method for resolving small samples and nonlinearities is SVM. The advantages are demonstrated by a maximum-dimensional design identification and other issues. Simply put, the goal of using the SVM approach is to identify a better classification hyperplane

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**
that completely separates the two types of data at the highest intervals. Regardless of whether the problem is two-type or multi-classification, the SVM has an optimal learning outcome. Originally, the 2-type problem was solved using SVM techniques. The following describes some of the basic rules in the second classification. An instance set for training is ={x_1⋯x_n},X∈R^d. The label for the equivalent type is {y_1⋯y_n},y_i∈{1,-1}. Assume that there are n instances and that the training instance feature vector's dimension is d.

5. PROPOSED WORK

In this proposed work the features from the pre-trained CNN namely AlexNet and GoogleNet are extracted. Following that, the deep features are input to SVM for the appropriate classification of musical instrument classes. The size of the spectrograms is changed from 510x350 to 28x28. In the initial maximum pooling, the feature map and image are both 14x14 in size. In the second convolution layer are 50 feature maps that were acquired. The picture size remains the same at 7x7 for both the feature map and the image. The Pre-Trained model's output is utilized as input for SVM model, which classifies the musical instruments sound. Each instrument is given a class by the system, and each class compares the other classes.

The spectrograms are resized from 510x350 into 224 x 224 and the stride 4 x 4 given as input for AlexNet. The input layer has 3 feature maps. The first is the convolution layer, which has 96 feature maps and an activation function. The input image is 55 by 55 pixels, with an 11 x 11 filter and a 4 x 4 stride. The first max pooling keeps the feature map at 96, the image at 27 x 27, the filter at 3 x 3, and the stride at 2 x 2. The 256 feature map has been obtained in the second convolution layer. The filter size is 5 x 5, the stride is 1 x 1, and the image size is 23 x 23. The feature map is unaltered in the second pooling. The filter size shrinks to 3 × 3, the stride is 2 x 2, and the feature map is 11 x 11. The feature map is extended to 384 in the third convolution layer. The picture is 9 x 9, the filter is 3 x 3, and the stride is 1 x 1. The fourth convolution layer's feature is unaltered; the image is 7 x 7 and the filter is 3 x 3 with a 1 x 1 stride. The fifth convolution layer has a feature map of 256, a 5 x 5 image, a 3 x 3 filter, and a stride of 1 x 1. The third max pooling of the feature map is still 2 x 2, 3 x 3 for the feature and 2 x 2 for the stride. The confusion matrix is used to calculate the Precision, Recall, F-Score, and Accuracy. 300 of the samples were used for testing and 1200 were used for training in the proposed system.

6. EXPERIMENTAL RESULTS

6.1 DATASET

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**

The IRMAS database and the NSynth dataset were used to extract musical instrument sound data with durations ranging from one second to two minutes. 16 different musical instruments from four different instrument families, including string, woodwind, keyboard, and brass, were used to create 1500 music samples. To classify the musical instruments, 20% of the data were used for testing, whereas 80% of the musical samples were used for training.

6.2 PERFORMANCE MEASURES

The performance measures of classification of musical instruments sound using SVM with the AlexNet and GoogleNet features. Table 1 shows the AlexNet with SVM performance and Table 2 depicts the GoogleNet with SVM performance of all fifteen musical instruments. Fig. 4 and 5 represents the accuracy of AlexNet with SVM and GoogleNet with SVM.

Table 1. Classification of AlexNet with SVM

Class	Precision	Recall
Banjo	97.22	94.59
Bassclarinet	93.22	94.82
Bassoon	92.10	93.33
Cello	96.96	94.11
Clarinet	93.75	95.74
Contrabassoon	94	94.94
Flute	97.87	97.87
Frenchhorn	95.79	97.43
Guitar	99.13	97.45
Mandolin	94.20	92.85
Saxophone	92.59	94.93
Trombone	95.77	94.44
Trumpet	92.30	90.56
Tuba	94.59	92.10
Violin	94.79	96.80

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**

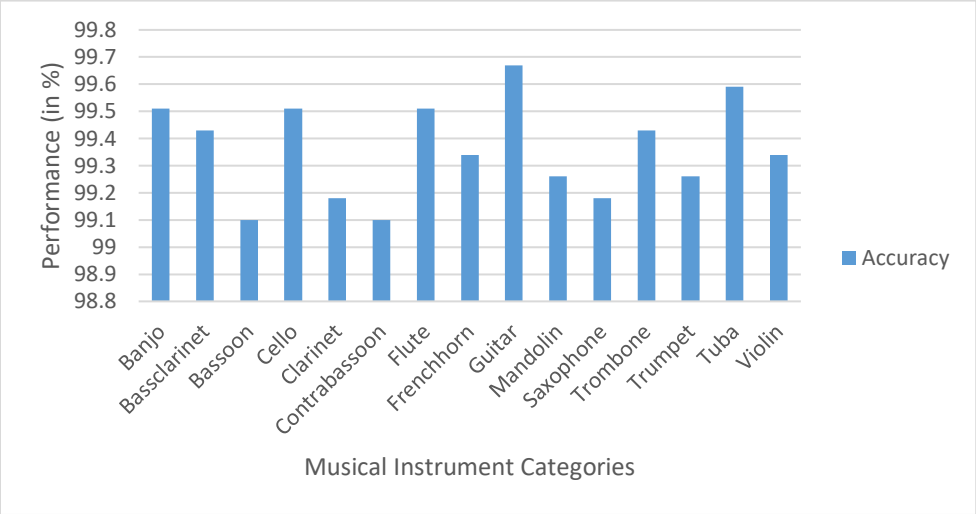

Fig. 4 Performance of AlexNet with SVM

Table 2. Classification of GoogleNet with SVM

Classes	Precision	Recall
Banjo	83.54	89.18
Bass Clarinet	80.35	77.58
Bassoon	85.71	80
Cello	78.04	95.52
Clarinet	83.33	85.10
Contrabassoon	88.88	88.88
Flute	88.65	88.65
French Horn	92.52	84.61
Guitar	90	76.27
Mandolin	75.28	95.71
Saxophone	81.17	87.34
Trombone	86.36	79.16
Trumpet	81.81	84.90
Tuba	66.66	84.21
Violin	87.67	68.08

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 10 (Oct -2025)
https://musikinbayern.com **DOI https://doi.org/10.15463/gfbm-mib-2025-482**

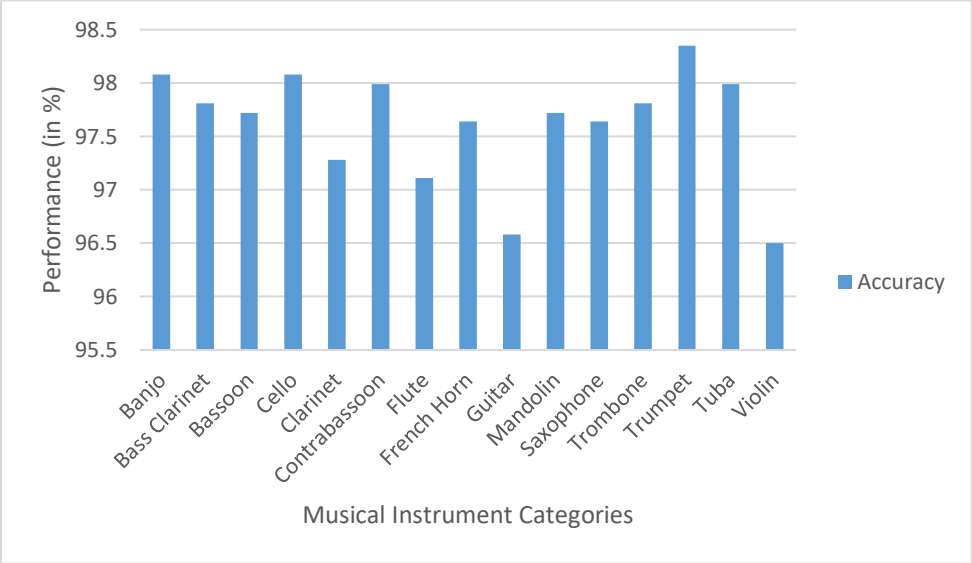

Fig. 5 Performance of GoogleNet with SVM

7. Performance Comparison with Existing Works

The performance of the proposed work is compared with some of the existing works. Table 3 shows the performance analysis of other works of various authors related to this paper.

Reference	Algorithms	Accuracy
Chandwadkar, D. M., & Sutaone, M. S. (2012) [10]	Sequential Minimal Optimization	97
E. Chaudary, S. Aziz, M. U. Khan and P. Gretschmann (2021) [11]	linear Support Vector Machine	94
N. Karunakaran and A. Arya (2018) [12]	hybrid classifier	90
Chakraborty, S.S., Parekh, R. (2018) [13]	cepstral coefficients	93
Anuz, H., Masum, A.K.M., Abujar, S., Hossain, S.A. (2021) [14]	MFCC with Support Vector Machine	97

Table 3 Performance Comparison of various works

8. CONCLUSION

This chapter has presented a set of DL based MISC models using AlexNet and GoogleNet models. The presented model undergoes pre-processing, feature extraction and classification. Once the

features are extracted using AlexNet and GoogleNet models, SVM and KNN models are applied for classification purposes. Among all the compared and different versions of proposed model, the GoogleNet-SVM model has outperformed the other by obtaining maximum precision of 96.94%, recall of 96.86% and accuracy of 97.22%.

9. REFERENCES

[1] Duda, R.O., Hart, P.E. and Stork, D.G., 2001. Pattern classification. John Wiley & Sons.

[2] Weihs, C., Ligges, U., Mörchen, F. and Müllensiefen, D., 2007. Classification in music research. Advances in Data Analysis and Classification, 1(3), pp.255-291.

[3] Scaringella, N., Zoia, G. and Mlynek, D., 2006. Automatic genre classification of music content: a survey. IEEE Signal Processing Magazine, 23(2), pp.133-141.

[4] Zölzer, U. ed., 2011. DAFX: digital audio effects. John Wiley & Sons.

[5] Malu, S., & Siddharthan, A. (2000). Acoustics of the Indian Drum. *arXiv: Mathematical Physics*.

[6] Gillet, O. and Richard, G., 2008. Transcription and separation of drum signals from polyphonic music. IEEE Transactions on Audio, Speech, and Language Processing, 16(3), pp.529-540.

[7] Yoshii, K., Goto, M. and Okuno, H.G., 2004, October. Automatic Drum Sound Description for Real-World Music Using Template Adaptation and Matching Methods. In ISMIR (pp. 184-191).

[8] Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E. and Torres-Carrasquillo, P.A., 2006. Support vector machines for speaker and language recognition. Computer Speech & Language, 20(2-3), pp.210-229.

[9] Aucouturier J & Pachet F. (2002). Representing Musical Genre: A State of Art, Journal of New Music Research, 83-93.

[10] Chandwadkar, D. M., & Sutaone, M. S. (2012). Role of features and classifiers on accuracy of identification of musical instruments. 2nd National Conference on Computational Intelligence and Signal Processing (CISP).

[11] E. Chaudary, S. Aziz, M. U. Khan and P. Gretschmann, "Music Genre Classification using Support Vector Machine and Empirical Mode Decomposition," 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), 2021, pp. 1-5, doi:

10.1109/MAJICC53071.2021.9526251.

[12] N. Karunakaran and A. Arya, "A Scalable Hybrid Classifier for Music Genre Classification using Machine Learning Concepts and Spark," 2018 International Conference on Intelligent Autonomous Systems (ICoIAS), 2018, pp. 128-135, doi: 10.1109/ICoIAS.2018.8494161.

[13] Chakraborty, S.S., Parekh, R. (2018). Improved Musical Instrument Classification Using Cepstral Coefficients and Neural Networks. In: Mandal, J., Mukhopadhyay, S., Dutta, P., Dasgupta, K. (eds) Methodologies and Application Issues of Contemporary Computing Framework.

[14] Anuz, H., Masum, A.K.M., Abujar, S., Hossain, S.A. (2021). Musical Instrument Classification Based on Machine Learning Algorithm. In: Tavares, J.M.R.S., Chakrabarti, S., Bhattacharya, A., Ghatak, S. (eds) Emerging Technologies in Data Mining and Information Security. Lecture Notes in Networks and Systems, vol 164.

[15] Prabavathy, S., Rathikarani, V., Dhanalakshmi, P. (2022). Musical Instrument Sound Classification Using GoogleNet with SVM and kNN Model. In: Chen, J.IZ., Tavares, J.M.R.S., Iliyasu, A.M., Du, KL. (eds) Second International Conference on Image Processing and Capsule Networks. ICIPCN 2021. Lecture Notes in Networks and Systems, vol 300. Springer, Cham.