RESEARCH PAPER

Musical Instruments Sound Classification using Spectrograms with Convolutional Neural Network

S. Prabavathy[1], G. Sophana[2] R. Selvalakshmi[3], G.Sangeetha[4]
[1, 3]Assistant Professor, [2]Assistant Professor and Head
[1, 2, 3] Department of Computer Science
[1, 2, 3] Kamaraj College (Autonomous), Thoothukudi, Tamilnadu, India.
[4]Arulmigu Subramania Swamy Arts and Science College, Vilathikulam, Tamilnadu, India.

Abstract-Many different instruments can interact to create music. In essence, people can tell which musical instrument was played, but it is challenging for a computer to do so on its own. Furthermore, different instruments produce varied sounds depending on timbre, playing style, and other characteristics, which makes identifying music more difficult. Understanding the type of instruments used in the musical audio sample is crucial when applying the Musical Instruments Sound Classification program. In this paper, ten different types of musical instruments from two musical instruments family were taken for research. The deep learning algorithm namely Convolutional Neural Network is used for the classification of musical instruments sound. In this paper, the work yields good result when compared with the existing algorithms.

Keywords: *Musical Instruments Sound, digital representation, CNN.*
How to cite this article: S. Prabavathy1 G. Sophana2 R. Selvalakshmi3 G.Sangeetha4. Musical Instruments Sound Classification using Spectrograms with Convolutional Neural Network. Int J Drug Deliv Technol. 2026;16(82s):620-625. DOI: 10.25258/ijddt.16.82s.69.

1. INTRODUCTION

1.1 MUSICAL INSTRUMENTS SOUND

Music is available in digital form. Several users have begun to buy music as audio files than physical medium, or acquiring music online. The digital distribution as well as application of music has resulted in massive opportunities for automatic process. Searching is done by examining the related metadata and audio content, for classifying and organizing the music in diverse forms. It becomes highly advantageous for listeners, enabling them to save their personal collections, and artists as well as music vendors, find it economical also as it is lesser in terms of cost but equal in terms of quality.

Musical Instruments Sound analysis is carried out in different formats. Digital audio techniques have now achieved domination in audio delivery with CD players, internet radio, mp3 players and iPods being the systems of choice in many cases [1]. In the research field, musical audio classification is emerging as a significant one because of the necessity of classifying and to categorize the music data automatically. The term "Audio" refers to a variety of audio signals, including voice, music, and more typical sound signals, as well as their combinations. Thousands of sound recordings are convenient for file systems or multimedia libraries. However, music is handled as a dense collection of bytes since it is attached to the most basic variables, such as file type, sampling rate, name, etc. Waveforms of digital music offer useful information that facilitates effective data comparison and classification.

1.2 DIGITAL REPRESENTATION OF AUDIO

Digital processing is now getting preference for handling Music and Audio. This is because audio applications and systems are mostly digital in nature. For processing a music signal the acoustic variations are to be represented in digital format. Representing the sound of musical instruments are not unique.

The four basic perceptual attributes of musical audio are

Pitch / Fundamental frequency

Pitch is a sound perceptual property that makes it feasible to order on a frequency-related scale. Pitch is the identification of sounds as “higher” and “lower” within the sense related to musical melodies. It is often determined merely in audio that have a frequency that it stable and clear enough to differentiate from noise.

Loudness / Amplitude

Volume of a music audio wave is determined by its amplitude, the larger for louder sound and the smaller for softer sound. Sound wave’s amplitude is set by the vibration that transmits energy to the medium.

Duration

Pitch of the musical audio wave marks a time length or duration that is controlled to make rhythm. An audio wave continues its path until we stop it.

Timbre

Timbre is also called as timber, the quality of audio vibrations formed by the sound wave tone. In music timbre is the attribute tone colour of a musical instrument sound or voice that occurs due to strengthening by individual instruments or singers of different overtones, or harmonics of a primary pitch.

1.3 TYPES OF MUSICAL INSTRUMENTS

There are many types of musical instruments available in our world. In this work the two instruments family is taken for research. They are Brass instrument and Woodwind instrument [18].

Brass Instrument

Brass instruments are made with brass or various metals to make sound by blowing air inside. Musician's mouth has to buzz, as to make a "raspberry" noise adjacent to the mouthpiece. Then air vibrates within the instrument, which makes a sound. Trumpet, Cornet, Trombone, French Horn, Tuba, and Bugle are a few brass instruments. Trumpet, Trombone, French horn and Tuba sounds are classified in this work.

Woodwind Instrument

Woodwind instruments make sound when air is blown inside. The sound comes from the instrument when the air vibrates inside. Woodwind instruments consist of Flute, Bass Clarinet, Clarinet, Saxophone, Contrabassoon, Bassoon, Piccolo, Recorder, and Oboe. For example, a flute made up of metal does not generate a substantially diverse sound from the one made of wood, for vibrations in a column of air in an instrument. The features of timbre with wind instruments are based on alternate factors, which are pointed with a length as well as shape of the musical tube. Length of a tube computes a pitch and however influences the timbre the piccolo, which is of partial size of a flute, with a shriller sound. Flute, Clarinet, Bass Clarinet, Saxophone, Contrabassoon, and Bassoon sounds are classified in this work.

1.4 MUSICAL INSTRUMENTS SOUND CLASSIFICATION

Music analysis has been an extremely active research topic. Each style of instrument has its own sound quality and is associated with particular forms of music. Experienced musical listeners can easily recognize the musical instruments that are present in a music recording, although identification varies in accuracy with the importance of an instrument (in the music), number of instruments played, familiarity, etc., Classification of music signal by humans is an implicit task, but a little challenging for computer systems. The use of digital audio signal processing includes a series of processes, for the digitalized audio signal compression, audio effect production and audio classification. The great importance of multimedia content management in audio classification and segmentation can be observed in today's world. The major issues are experienced in audio and visual data handling. It is occupied by audio classification with important applications in broad fields [2]. In this paper, the signals of musical instruments sound are transformed into spectrograms and then fed into the Convolutional Neural Network to classify the musical instruments sound categories. The Fig. 1 presents the block diagram of the proposed work.

Fig. 1. Block diagram of the proposed work

2. LITERATURE REVIEW

The author in [3] applied a hierarchical model for audio categorization and retrieving according to the audio investigation. It is limited to three stages. The term "coarse-level audio classification" refers to the segmentation process that divides recordings into speech, music, alternate ecological music, and silence. This process relies on both statistical and morphological analysis of temporal curves for short-term audio signal characteristics. Second, ecological noises are separated into various classes, such as the sounds of birds, rain, applause, and so forth. Time-frequency analysis of audio data and the use of the Hidden Markov Model (HMM) for classification are the fundamental components of categorization.

Both monophonic and polyphonic musical contents were classified by the author [4]; in the former case, a single instrument is played, while in the latter case, multiple instruments are played simultaneously [5]. While the classification of instruments in monophonic contents is successful, in polyphonic, it is challenging to analyze the annotation of musical signals [6]. The promising issues in MIC are the variation in timbre and functioning of the musical signal combined with the perceptual affinity of additional musical instruments and the superposition of various instruments under the frequency and time [7]. Additionally, choosing the best feature set for the applied classification method is another challenge in dividing the musical signal. Consequently, the deep learning approach improves the model's ability to extract features from real musical signals and provides benchmark results for analyzing the signal. The problems of structured coding, automated annotation of musical instruments, and retrieving musical instruments from a database system are all resolved by automated musical classification [8].

In [9], a system for classifying musical blocks into seven classes was proposed, and the problem of frequent and typical audio content-based retrieval was reported. Known as feature vectors, it is also used with MFCC and LPC. Compared to performance validation, a higher accuracy has been achieved. For ecological sound analysis, [10] implemented a comprehensive comparison study of Artificial Neural Network (ANN), Learning Vector Quantization (VQ), and Dynamic Time Warping (DTW) classifiers combined with static or non-static feature extraction. Maximum recognition using MFCC or continuous wavelet transform in conjunction with DTW has been demonstrated in the final results.

Using test sets with identical classification models and general training, the author [11] reported the differentiation achieved by different characteristics. The test-organized database includes music and audio in 13 different languages worldwide. At this stage, a Gaussian Mixture Model (GMM) is used to identify the feature space distributions. Amplitude, cepstra, pitch, and zero crossings are the four types of characteristics used to calculate performance. To improve the function, a feature's derivative is first applied and referred to. Applying cepstra and delta

cepstra, which have comparable equal error rates (EER), results in the best performance. Both normal amplitude and delta amplitude have successively followed it. However, it is made up of the fewest possible complications. With 0% crossing and maximum EER, the pitch and delta pitch produced superior EER.

A number of recent developments have skillfully invoked the advancements of the contemporary deep learning methodology. It is employed as an optimization technique, which is more important. When Rectified Linear Units (ReLUs) were used instead of sigmoid functions, the Deep Neural Network's training speed rose quickly [12]. Consequently, speech recognition [14] and picture recognition [13] have been introduced [15]. A major hardware advancement is an alternative and important change. Large-scale visual image classification is thought to have been pioneered by [13] with the use of Graphics Processing Units (GPUs) in parallel computing. The deep learning methodology and the basic neural network unit have been developed in this model.

For MISC models, the author [16] used hierarchical topologies. In order to learn the abstractions of feature implications, learning methodologies could employ hierarchical topologies by stacking non-linear network layers. The learning strategy can be used to map the necessary procedures to data structures for the applied computation and to identify contextual dependencies. CNNs have demonstrated greater efficiency in these connections for MISC applications because to their effective performance under data fluctuations. An audio sample must be manually analyzed for engineer implications before being fed into a deep learning network. While hand-based features are useful for large-scale problems, their performance is limited and their results are minimal [17]. Consequently, there is the possibility of studying feature representations with particular trade-offs, followed by learning strategies like CNN that necessitate larger datasets, the highest processing power, and a thorough understanding of machine learning to effectively train these techniques. In any case, the most impressive results for deep learning models have been shown by the present implementations for image and audio-based operations.

3. CONVOLUTIONAL NEURAL NETWORK

Deep Learning is a subfield of machine learning. Convolutional Neural Network (CNN) is a collection of deep learning which is mainly used for audio classification, image classification, image recognition, video labelling, natural language processing, medical image analysis etc. Convolutional Neural Networks consists of output layer, input layer and several hidden layers. The hidden layers consist of many convolution layers. The Convolutional Neural Network model is designed with a simple stack of multiple convolutional layer followed by ReLU activation function proceeded by max pooling layers. More significantly, it is used as an optimization method [21, 22]. The training speed of Deep Neural Network was increased in a rapid manner using Rectified Linear Units (ReLUs) than applying the sigmoid functions. As a result, speech recognition [14] and image recognition has been presented [15]. An alternate and significant modification is a massive hardware development. Parallel computing on Graphics Processing Units (GPUs) deployed by [12] is considered to be the pioneer of large-scale visual image classification. In this model, the fundamental units of neural network have been defined along with deep learning approach. Fig. 2 illustrates the architecture of Convolutional Neural Network.

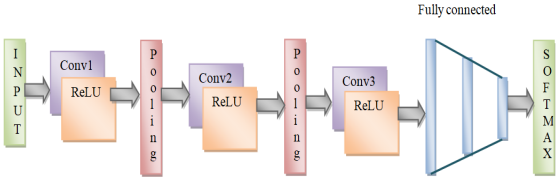

Fig. 2. Architecture of Convolution Neural Network

Steps of Convolutional Neural Network

1. Provide input image inside convolutional layer.
2. Choose parameters, apply filters along with strides, padding if necessary.
3. Perform convolution on the image and use ReLU activation to the matrix.
4. Render pooling to decrease dimensionality size.
5. Contribute as many convolutional layers until contented.
6. Flatten the outcome and contribute into a fully connected (FC) layer.
7. Outproduce the class handling an activation function.

4. PROPOSED WORK

4.1 CONVOLUTIONAL NEURAL NETWORK

In the proposed framework, the CNN structure is designed with a simple stack of multiple convolution layers proceeded by ReLU activation function followed by max pooling layers. The proposed CNN model consists of input layer, three convolutional layers, three pooling layers, two FC layers and the output layer.

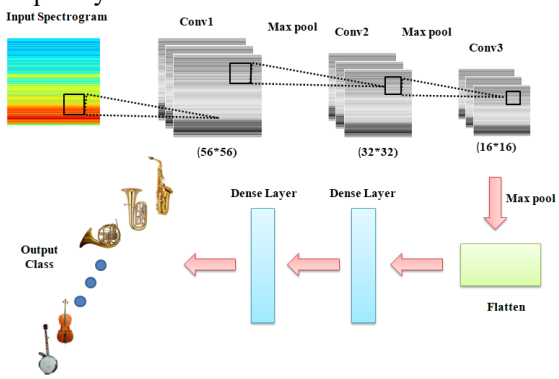

Fig. 3. Proposed Framework of MISC using CNN

Here the model consists of 10 output classes namely Banjo, Bass Clarinet, Bassoon, Cello, Clarinet, Contrabassoon, Flute, French Horn, Guitar, Mandolin, Piano, Saxophone, Trombone, Trumpet, Tuba and Violin. The spectrogram images are generated from the input musical audio signal which is given as input to the Convolutional Neural Network model for the classification of musical instruments sound. The size of the spectrogram image is 64 x 64. Fig.3. shows the proposed framework of MISC using Convolutional Neural Network.

4.2 SPECTROGRAM

Spectrogram is an image of music / sound. Audio signals are applied to create a spectrogram which is also called as sonographs, voicegrams or voiceprints. A spectrogram shows the frequencies that make up the sound, from least to most, and how they change after some time, from left to right, frequencies on vertical axis, and the time on horizontal axis. A spectrogram is usually depicted as a heat map, i.e., as an image with the intensity shown by varying the color or brightness. The spectrogram image is a visual implication of frequency spectrum, machine learning methods are employed for classifying the spectrogram images.

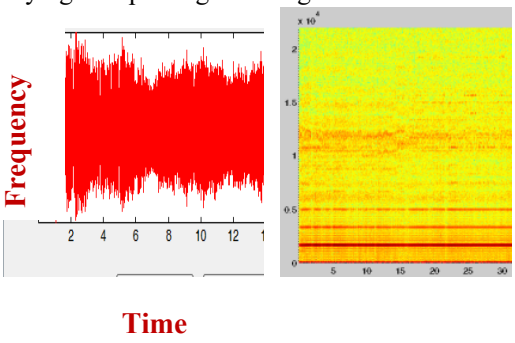

Fig. 4. Spectrogram of the Input Music Signal

The music signal is usually shown as a heat map as a picture. Fig. 4 shows the music signal and spectrogram image of Flute.

5. EXPERIMENTAL RESULTS

5.1 DATASETS

The music data are obtained from IRMAS online musical database. 1000 musical audio clips were collected such as 100 audio clips of French Horn, Trombone, Trumpet, Tuba, Bass Clarinet, Bassoon, Clarinet, Contrabassoon, Flute and Saxophone respectively. Each music clip consists of music data ranging from 1 second to 10 seconds duration and the music is sampled at 8 kHz and encoded by 16-bit. 750 musical data samples were used for training and 250 for testing. The input music clips were converted into spectrograms before classification.

5.2. Performance Evaluation for CNN models

This model consists of three convolutional layers followed by pooling layer, FC layer and exit with the output layer. The spectrogram is resized into 64 x 64 before classification and the size of the kernel is 5 x 5. In the first convolutional layer, 80 filters have been convolved to the input image and it yields 630 parameters, Maximum pooling and Rectified Linear Unit is used for the initial convolutional layer. In the second convolutional layer, 120 filters, have been utilized which yields 4096 parameters. In the third convolutional layer, 120 filters have been used and they yield 6144 parameters. Maximum pooling and ReLU have been utilized for all the convolutional Layers. Finally, these are flattened and 18015 features are obtained and it is fed into the dense layer. The same procedure is repeated for 3 seconds, 5 seconds and 10 seconds duration musical dataset. For testing, the spectrogram images of size 64 x 64 are given as input for the classification of musical instruments sound. The prediction matrix is given to the sixteen categories of sound of the musical instruments. The performance is calculated based on the confusion matrix on the predicted value. Performance metrics namely Precision, Recall, F-Score and Accuracy are used. Table 1. shows the performance evaluations for CNN model.

Table 1 Performance Evaluations for CNN Model

Classes	Precision (in %)	Recall (in %)	F-Score (in %)
French horn	83.80	90.20	86.88
Trombone	97.80	96.70	97.25
Trumpet	95.60	95.60	95.6
Tuba	96.20	97.50	96.85
Bassoon	95.70	96.40	96.05
Bass clarinet	98.10	97.80	97.95
Clarinet	81.40	92.30	86.51
Contrabassoon	95.90	96.20	96.05
Flute	93.70	94.00	93.85
Saxophone	98.30	96.40	97.34

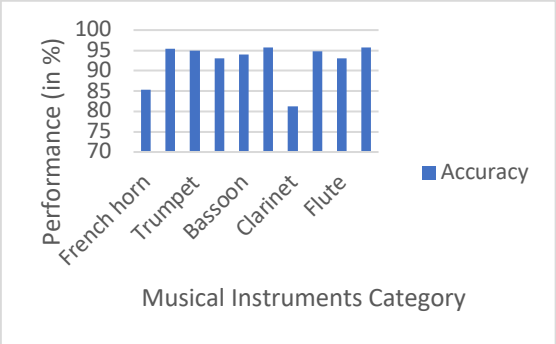

Fig. 5 Accuracy of CNN Model

The accuracy of this work shows in Fig. 5. The performance of CNN model for various durations of music data is shown in Fig. 6. In this work 10 seconds duration yields highest accuracy when compared to others.

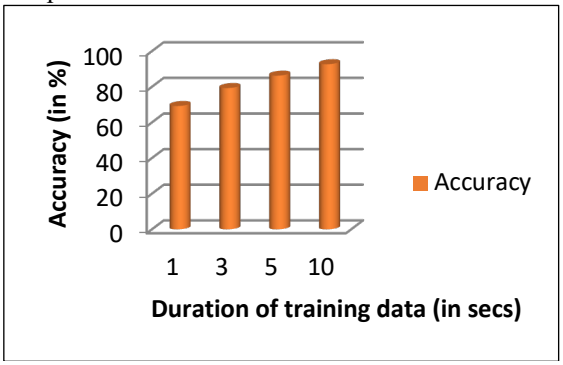

Fig. 6. Performance of CNN Model

6. COMPARISON WITH EXISTING WORKS

The proposed work is compared with various existing works. Table 2 shows the performance comparison.

Table 2. Comparison of existing algorithms

Reference	Algorithms	Accuracy
Tamanna, et al., (2024) [19]	CNN	85%
Sivalingam, et al., (2025) [20]	CNN	78.37
Chandrashekar, et al., (2025) [23]	SVM	75
Patil, et al., (2025) [24]	CNN	90
Prabavathy S and Selvalakshmi R [25]	SVM	95

7. CONCLUSION

In this paper, the classifications of musical instruments sound have been categorized. An effective deep learning based CNN model is developed for classifying the musical instruments sound effectively. To classify the musical signal, the deep learning method has been implemented using the spectrograms. The experimental results ensured the performance of the CNN model with the maximum precision of 93.65%, recall of 95.31% and accuracy of 92.30%. The yield results are good when compared with the existing models.

REFERENCES

[1] Ian Mc Loughlin. (2009). Applied Speech and Audio Processing: With MATLAB Examples, *Cambridge University Press*.

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

[3] Wang, Y., Yun, X., Zhang, Y., Chen, L. &Zang, T. (2017). Rethinking Robust and Accurate Application Protocol Identification. *Computer Networks*, *129*, 64-78.

[4] Fu, Z., Lu, G., Ting, K.M. & Zhang, D. (2011). A Survey of Audio-based Music Classification and Annotation. *IEEE Transactions on Multimedia, 13(2),* 303-319.

[5] Herrera-Boyer, P., Peeters, G. & Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research*, *32*(1), 3-21.

[6] Choi, K., Fazekas, G., Cho, K. & Sandler, M. (2017). A Tutorial on Deep Learning for Music Information Retrieval. 16.

[7] Newton, M.J. & Smith, L.S. (2012) A Neurally Inspired Musical Instrument Classification System based upon the Sound Onset. *The Journal of the Acoustical Society of America*, *131*(6), 4785-4798.

[8] Muhury, S., Neogi, G., Debnath, P. &Ghosh Dastidar, J. (2016). Design of a voice-based system by recognizing speech using MFCC. *Computational Science and Engineering Proceedings of the International Conference on Computational Science and Engineering (ICCSE2016), CRC Press,* 77–80.

[9] Li, D., Dimitrova, N., Li, M. & Sethi, I.K. (2003). Multimedia Content Processing through Cross-Modal Association. In *Proceedings of the Eleventh ACM International Conference on Multimedia,* 604-611.

[10] Cowling, M. & Sitte, R. (2003). Comparison of Techniques for Environmental Sound Recognition. *Pattern Recognition Letters*, *24*(15), 2895-2907.

[11] Nosratighods, M., Ambikairajah, E., Epps, J. & Carey, M.J. (2010). A Segment Selection Technique for Speaker Verification. *Speech Communication*, *52* (9), 753-761.

[12] Glorot, X., Bordes, A. & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics,* 315-323.

[13] Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *InAdvances in Neural Information Processing Systems,* 1097-1105.

[14] Dahl, G.E., Sainath, T.N. & Hinton, G.E. (2013). Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* 8609-8613.

[15] Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q.V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J. & Hinton, G.E. (2013). On rectified Linear Units for Speech Processing. *In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* 3517-3521.

[16] Dieleman, S. (2015). *Learning Feature Hierarchies for Musical Audio Signals.* [Doctoral Dissertation]. Ghent University, Faculty of Engineering and Architecture, Ghent, Belgium.

[17] Humphrey, E. J., Bello, J.P. & LeCun, Y. (2012). Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. *In ISMIR,* 403-408.

[18] Aurchana, P. K., & Prabavathy, S. (2021). Musical instruments sound classification using GMM. *London Journal of Social Sciences*, (1), 14-25.

[19] Tamanna, Sheeban E and Ezhan, Mohammed and Mahesh, R and Shetter, (2024), Anupama and Parameshachari, B D and Sunil Kumar, D S and Puttegowda, Kiran, "Musical Instrument Classification Using Deep Learning CNN Models," *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India*, pp. 01-07.

[20] Sivalingam, Padmesh and J, Aamith Kishore T and P, Sri Krishna and B, Yaswanth Reddy and S, Ragav and R., Lekshmi C., "Robust CNN-based Musical Instrument Recognition with Enhanced Feature Learning," *2025 International Conference on Inventive Computation Technologies (ICICT)*, Kirtipur, Nepal, 2025, pp. 851-857.

[21] Liu, Y. (2025). Deep Neural Network and SVM-Based Multiclass Musical Instruments Classification Using MFCC, STFT, and Related Acoustic Features. *Informatica*, *49* (33).

[22] Bin C., (2025) "A Sound Classification Method of Traditional Chinese Musical Instruments Based on MEL Spectrogram and Convolutional Neural Networks," *2025 IEEE 12th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Chongqing, China, pp. 369-373.

[23] Chandrashekar, Harshith and K T, Swetha and V, Ambika and Puttegowda, Kiran and Kumara, Bharath and R, Paramesha and B D, Parameshachari, (2025) "Music Instruments Classification Using Signal Processing and Machine Learning," *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*, BENGALURU, India, pp. 1-7.

[24] Patil N., Kudari M., Nandeppanavar A. S. and Thotad P. N., (2025) "Deep Learning for Automatic Detection of Musical Instruments in Audio Signals," *2025 5th Asian Conference on Innovation in Technology (ASIANCON)*, PIMPRI, India, pp. 1-6.

[25] S. Prabavathy, R. Selvalakshmi" Sound Classification of Musical Instruments with Sonogram Features using Machine Learning Algorithms" Musik In Bayern, Vol. 91, Issue 4, April 2026, pp1-22.