Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00281-x

Original Article

Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models

Sushant Sharma [1], Rushali R. Katkar [1], Sushama A. Shirke [1], Shreya Jare [1]

1. *Department of Computer Engineering, Army Institute of Technology, Pune, IND*

Received: July 08, 2026 | Review began: July 29, 2026 | Review ended: August 23, 2026 | Published: September 16, 2026

Abstract

Background

Environmental Sound Classification helps smart systems understand sounds happening in real-life situations. It is used in areas like watching over places, making cities smarter, keeping an eye on health, and better understanding media content.

Objective

This study compares three traditional machine learning classifiers - K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest (RF) - with a Convolutional Neural Network (CNN) for classifying environmental sounds using the UrbanSound8K benchmark.

Methods

For the classical models, four groups of handcrafted features - Mel-Frequency Cepstral Coefficients, Chroma, spectral features, and Tonnetz - were extracted from the audio recordings. The CNN model used log-Mel spectrograms as its input. All models were tested using the 10-fold cross-validation method provided by UrbanSound8K. In each round of testing, preprocessing and scaling were done only on the training data and then applied to the test part. The results were measured using accuracy, precision, recall, macro F1-score, confidence intervals, and paired statistical tests.

Results

Among the classical approaches, RF achieved the highest macro F1-score (65.72%), followed by linear SVM (62.04%) and KNN (56.26%). The proposed CNN obtained a macro F1-score of approximately 70.42% with an accuracy of approximately 69.35%, outperforming all classical baselines. Both the paired Student's t-test and Wilcoxon signed-rank test indicated statistically significant differences between the CNN and each classical baseline ($p < 0.05$).

Conclusion

The results demonstrate that CNNs provide superior environmental sound classification performance by learning discriminative spectral-temporal representations directly from log-Mel spectrograms, whereas RF remains a computationally efficient alternative for resource-constrained deployments. A graphical user interface-based prototype is presented as a proof-of-concept deployment framework illustrating practical inference rather than as a separately evaluated scientific contribution.

Categories: AI applications, Machine Learning (ML), Deep Learning

Keywords: environmental sound classification, urbansound8k, machine learning, audio classification, feature extraction, convolutional neural network, audio signal processing, deep learning

Introduction

Environmental sound classification (ESC) has emerged as a relevant task within machine learning and audio signal processing because of its ability to enable intelligent systems to interpret surrounding acoustic environments [1]. Unlike speech recognition, which focuses primarily on linguistic content, ESC aims to identify non-speech audio events such as sirens, dog barks, engine noise, drilling, and gunshots. Accurate recognition of such sounds provides valuable contextual information for numerous real-world applications,

including intelligent surveillance, smart city monitoring, healthcare assistance, autonomous systems, multimedia indexing, and human-computer interaction. As urban environments continue to generate increasingly diverse acoustic information, reliable ESC systems have become an essential component of context-aware artificial intelligence.

Environmental sound recognition presents several challenges due to the complex nature of real-world acoustic environments. Background noise, overlapping sound events, variations in recording devices, and significant intra-class variability often reduce classification performance. Furthermore, many environmental sounds exhibit similar spectral characteristics, making discrimination between acoustically related classes particularly difficult. These challenges require classification models capable of learning robust and discriminative representations from noisy and highly variable audio signals.

Traditional ESC systems usually use manually created acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, spectral descriptors, and Tonnetz representations [2]. These features help capture different types of sound information, such as timbre, harmony, and spectrum, while staying computationally efficient and easy to understand. Because of this, older machine learning methods such as K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forests (RF) are still useful for applications with limited resources and on edge devices. However, their performance depends a lot on the quality of the manually designed features, which might not fully capture the complex time and frequency patterns found in environmental sounds.

Deep learning approaches have improved ESC by learning task-relevant features directly from time-frequency representations such as log-Mel spectrograms [3]. CNNs can learn hierarchical representations of spectral and temporal patterns [2,3] while reducing the dependence on manually engineered features. More recent ESC studies have explored attention mechanisms, hybrid architectures, knowledge distillation, and lightweight model designs to improve representation learning and computational efficiency [4-7]. However, increasing model complexity can also raise computational and training requirements, making the balance between classification performance and deployment efficiency an important consideration for practical ESC systems.

Although numerous studies have investigated either classical machine learning or deep learning approaches independently, comparatively fewer studies provide a systematic evaluation of both paradigms using identical datasets, standardized evaluation protocols, and consistent performance metrics. In addition, many published studies primarily report average accuracy while providing limited analysis of fold-wise variability, confidence intervals, or statistical significance of the observed improvements. Such analyses are important for determining whether reported performance gains represent genuine improvements rather than variations arising from cross-validation splits.

To tackle these issues, this study compares classical machine learning and deep learning methods for classifying environmental sounds using the UrbanSound8K dataset [1]. KNN, SVM, and RF are tested with several different types of manually created acoustic features, while CNN is trained using log-Mel spectrograms, following the same official 10-fold cross-validation process. The models' performance is measured using accuracy, precision, recall, macro F1-score, confidence intervals, and statistical tests to ensure fair comparison. Additionally, a GUI is created as a basic example of how these models can be used in real-world applications.

The main contributions of this study are as follows:

Complete model comparison: A thorough evaluation of KNN, linear SVM, RF, and CNN models using the official UrbanSound8K 10-fold cross-validation method.

Acoustic feature analysis: A comparative investigation of multiple handcrafted acoustic feature groups, including MFCCs, Chroma, Spectral, and Tonnetz descriptors, to determine their effectiveness for environmental sound classification.

Rigorous experimental evaluation: Performance is evaluated using macro F1-score, accuracy, precision, recall, fold-wise statistics, confidence intervals, and paired statistical significance tests to ensure reliable comparison between models.

Deployment demonstration: A GUI-based prototype integrates audio extraction, feature processing, model inference, and semantic interpretation to illustrate the practical deployment of environmental sound classification systems rather than introducing a separate scientific contribution.

Materials And Methods

Dataset

The experiments used the UrbanSound8K dataset [1], which is a standard set of sounds used to classify different types of environmental noises. It has 8,732 audio clips that have been labelled by people, and these clips fall into 10 different categories: air conditioner, car horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren, and street music. Each audio clip is no longer than 4 s and is stored as a single-channel WAV file.

UrbanSound8K comes with a standard way to split the data into 10 parts, which is used to test how well models perform. To make sure results are consistent with earlier work and to avoid sharing information between training and testing, the splits were used as they were. For each testing round, one part was used to test the model, while the other nine parts were used to train it. This process was done for all 10 parts, so each part was tested once. The results from all the tests were then averaged to give a final performance score.

Experimental workflow

To make sure the models are compared fairly and to prevent any information from leaking between them, all the steps for preparing the data were done separately in each group of data used for testing. For each of the 10 rounds in the official cross-validation process, this was done:
One specific group from the UrbanSound8K dataset was chosen as the test group, and the other nine groups were used for training.
Features that describe the sound were created separately for each audio file.

For the traditional machine learning models, a process called standardisation was used. This was done only on the training data, and the values calculated from that data were used to adjust the test data without redoing the process.

Each model was trained only on the standardized training data and then tested on the test data that it had never seen before.

For each round, results like accuracy, precision, recall, and the macro F1-score were noted down.

After all 10 rounds were completed, the overall results were combined by taking the average, the spread of results, and the 95% confidence interval across all rounds.

For the CNN model, visual representations of the sound called log-Mel spectrograms were made for all the audio files.

The predefined groups for cross-validation were kept the same during both training and testing. In each round, only the training groups were used to improve the model, while the test group was not used at all until the model was ready to make predictions.

Feature extraction

In environmental sound classification, feature extraction turns raw audio recordings into simpler forms that are easier for machine learning algorithms to work with. In this study, all audio files were made to have the same sampling rate of 22.05 kHz to make sure everything in the dataset is consistent. Audio clips that were shorter than one second were made longer by adding zeros before feature extraction, which helps in calculating acoustic features accurately. Feature computation was done using the Librosa audio analysis library [8]. Unless stated otherwise, the default settings of Librosa were used, which include an FFT size of 2048 samples and a hop length of 512 samples.

Four different types of features were used to describe the sounds in the environment: MFCCs, Chroma features, Spectral features, and Tonnetz features. These features capture different parts of the sound, such as its spectral, harmonic, and timbral qualities, giving a strong and complete description for classification.

MFCCs: These are widely used features that show the spectral makeup of an audio signal and also reflect how humans hear sound.

$$\mathrm{MFCC}_n = \sum_{m=1}^{M} \log(S_m) \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \tag{1}$$

where

S_m represents the energy associated with the m^{th} Mel filter,

M denotes the number of Mel filters,

n represents the cepstral coefficient index.

In this work, 13 MFCCs were extracted from each audio recording. To obtain a fixed-dimensional representation, the mean and standard deviation of each coefficient were calculated over the analysis frames, allowing the resulting vector to capture both average spectral characteristics and temporal variation.

Chroma Features: Chroma features describe how spectral energy is distributed across the twelve pitch classes and provide complementary harmonic information. Although environmental sounds generally contain less harmonic structure than speech or music, chroma descriptors can improve discrimination for sound classes containing tonal components. The temporal mean of each chroma coefficient was used to obtain a fixed-length feature vector.

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

Spectral Features: Several spectral descriptors were extracted to characterise the frequency-domain properties of the environmental sounds. Among these descriptors, the spectral centroid represents the frequency-weighted centre of the spectrum and provides an estimate of where the spectral energy is concentrated, as expressed by

$$C = \frac{\sum_{k=1}^{N} f_k X(k)}{\sum_{k=1}^{N} X(k)} (2)$$

where f_k denotes the frequency corresponding to the k^{th} spectral bin, $X(k)$ represents the magnitude spectrum, and N denotes the number of frequency bins.

Along with the spectral centroid, four additional descriptors were extracted: Zero-Crossing Rate, Root Mean Square Energy, Spectral Bandwidth, and Spectral Rolloff. Together, these features describe the temporal and spectral characteristics of the audio signal, including energy distribution, frequency concentration, spectral spread, and spectral decay. The temporal mean of each descriptor was used to obtain a fixed-dimensional feature representation suitable for classification.

Tonnetz Features: Tonnetz features were computed from the harmonic component of each audio signal following harmonic-percussive separation. These descriptors capture tonal relationships between frequency components and provide additional harmonic information that may assist in distinguishing acoustically similar environmental sounds. Similar to the chroma descriptors, the temporal mean of each Tonnetz coefficient was used as the final feature representation.

The extracted feature groups were evaluated both individually and in combination to determine their effectiveness for environmental sound classification. Experimental analysis showed that the combined MFCC + Chroma + Spectral representation produced the highest performance among the classical machine learning models and was consequently selected for comparison with the proposed CNN model. Before model training, all feature vectors were standardised using statistics computed exclusively from the training data for each cross-validation fold.

$$x' = \frac{x - \mu}{\sigma} (3)$$

where x denotes the original feature value, x' denotes the standardized feature value, μ represents the mean calculated from the training fold, and σ represents the corresponding standard deviation. The same transformation was subsequently applied to the corresponding test fold without recomputing these statistics, thereby preventing information leakage and ensuring an unbiased evaluation protocol.

The parameters used during feature extraction are summarized in Table *1*.

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

Parameter	Value
Sampling rate	22,050 Hz
Minimum audio length	1 s (zero-padded if shorter)
MFCCs	13
Mel filter banks (CNN)	64
FFT size	2048 samples (Librosa default)
Hop length	512 samples (Librosa default)
MFCC aggregation	Mean and standard deviation
Chroma aggregation	Mean
Spectral feature aggregation	Mean
Tonnetz aggregation	Mean

TABLE 1: Parameters used for feature extraction

FFT, Fast Fourier Transform; CNN, Convolutional Neural Network; MFCC, Mel-Frequency Cepstral Coefficients

Classical machine learning models

Three machine learning methods were chosen as baseline models: KNN, Linear SVM, and RF. These methods show different ways of learning: one uses similar examples, another focuses on decision boundaries, and the third combines multiple models. All models were trained and tested using the same 10-fold cross-validation method from the UrbanSound8K dataset. To make the comparison fair, all classifiers used the same training and testing data splits and the same standardised features.

The hyperparameters for each classifier were selected using commonly reported values and preliminary empirical evaluation. All experiments were implemented using the Scikit-learn library [9].

K-Nearest Neighbours (KNN)

KNN is a type of method that does not assume any specific form for the data. It works by looking at the closest data points to a given sample and then decides which class that sample belongs to based on what most of those nearby points are classified as. A neighbourhood size of $k = 5$ was employed. Because the feature vectors were standardised before classification, the default Euclidean distance (Minkowski distance with $p = 2p$) was used to identify the nearest neighbours. Uniform neighbour weighting was adopted, allowing each of the five neighbours to contribute equally to the final prediction.

Linear Support Vector Machine (SVM)

A linear kernel was used with the default regularisation parameter ($C = 1.0$). Class imbalance was addressed using balanced class weights, and probability estimates were enabled to facilitate confidence score computation. Multi-class classification was implemented using the one-versus-rest strategy provided by Scikit-learn. For the binary case, the optimisation problem is formulated as

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \quad (4)$$

subject to the constraints

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \,(5)$$

where

where $\mathbf{w}$ denotes the weight vector,

b denotes the bias term,

C represents the regularization parameter, and

ξi denotes the slack variables.

Random Forest (RF)

RF
is a technique that uses many decision trees together to make better predictions and prevent the model from becoming too focused on the training data. The final result is decided by the most common answer given by the trees.

$$\hat{y} = \mathbf{mode}\left\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_M(\mathbf{x})\right\} (6)$$

where

$Ti(x)$ denotes the prediction produced by the i^{th} decision tree,

M represents the total number of trees.

A RF consisting of 200 decision trees was employed. Balanced class weights were used to reduce the effect of class imbalance, while bootstrap sampling and random feature selection during tree construction increased model diversity and improved generalisation.

Comparison Protocol

To make sure all the classifiers worked the same way, we used the same features, steps to prepare the data, and ways to check how well they performed. We only used the training data to figure out how to standardise the features, and then we used those same rules on the test data. We measured how well the models worked using accuracy, precision, recall, and macro F1-score. The numbers we reported are the average results from all 10 different tests.

Convolutional Neural Network (CNN) architecture

CNN was developed using log-Mel spectrograms as input to model the temporal and spectral characteristics of environmental sounds. Unlike manually engineered acoustic features, CNNs can learn hierarchical representations directly from the input data, which can help distinguish acoustically similar sound classes.

Each audio recording was converted into a 64-band log-Mel spectrogram, which was either zero-padded or truncated to 173 time frames to produce a fixed-size input representation suitable for CNN processing. The resulting feature map was represented as a single-channel tensor of dimensions (64\times173\times164\). The logarithmic transformation applied to the Mel spectrogram can be expressed as

$$L = 10\log_{10}(M) \,(7)$$

where M represents the Mel spectrogram power and L denotes the resulting log-Mel spectrogram used as the CNN input.

The CNN structure has three convolutional blocks and a final classification part. The first, second, and third convolutional layers use 32, 64, and 128 filters respectively, each with a 3 × 3 kernel, same padding, and ReLU activation. After each convolutional layer, Batch Normalization is used to help the model train better and reach a good solution. Max-pooling is applied after the first two convolutional blocks to shrink the size of the data while keeping important features. Then, Global Average Pooling is used to turn the features into a smaller vector that can be used for classification. This vector goes into a fully connected layer with 128 neurons, followed by a dropout layer that randomly turns off some neurons with a rate of 0.30 to prevent the model from becoming too focused on certain parts. Finally, a Softmax layer gives the probabilities for the 10 different environmental sounds. The way each convolutional layer extracts features can be described with math.

$$Y(i,j) = \sum_m \sum_n X(i+m, j+n)\, K(m,n) + b \,(8)$$

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

where $\boldsymbol{X}$ denotes the input feature map, $\boldsymbol{K}$ represents the convolution kernel, $\boldsymbol{b}$ denotes the bias term, and $\boldsymbol{Y}$ denotes the resulting feature map.

The Rectified Linear Unit (ReLU) was applied after each convolutional layer to introduce non-linearity and is defined as

$$\mathbf{ReLU}(\boldsymbol{x}) = \mathbf{max}(\mathbf{0}, \boldsymbol{x}) \ (\mathbf{9})$$

The network was trained with the Adam optimizer using a learning rate of 0.001 and the Sparse Categorical Cross-Entropy loss function. The corresponding training loss is defined as

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \ (10)$$

where C represents the number of classes, y_i denotes the ground-truth class label, and $\hat{y}_i$ denotes the predicted probability generated by the Softmax layer.

The CNN was evaluated under the official 10-fold cross-validation protocol used for the classical machine learning models. For each iteration, nine predefined folds were used for training and the remaining fold was reserved as the independent test set. Within each training set, 10% of the samples were held out as a validation subset for early stopping and learning-rate scheduling. The performance metrics from the 10 folds were averaged to obtain the final evaluation results, allowing comparison with the classical baseline models.

To make the results repeatable, a fixed random seed of 42 was set for Python, NumPy, and TensorFlow. The CNN was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and up to 20 training epochs. In each training round, 10% of the training data was set aside for validation. Early stopping was used with a patience of 5, and ReduceLROnPlateau was applied with a factor of 0.5 and a patience of 3 to prevent overfitting and help the model converge. No data augmentation was used during training. Unless stated otherwise, all convolutional and dense layers used the default Glorot Uniform (Xavier) weight initializer from TensorFlow/Keras.

Table *2* summarizes the CNN architecture and training configuration.

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

Component	Configuration
Input	64 × 173 × 1 Log-Mel spectrogram
Conv Block 1	Conv2D (32, 3 × 3) + BatchNorm + MaxPooling
Conv Block 2	Conv2D (64, 3 × 3) + BatchNorm + MaxPooling
Conv Block 3	Conv2D (128, 3 × 3) + BatchNorm
Pooling	Global Average Pooling
Dense Layer	128 neurons, ReLU
Dropout	0.3
Output Layer	Softmax (10 classes)
Optimizer	Adam (learning rate = 0.001)
Loss Function	Sparse Categorical Cross-Entropy
Batch Size	32
Epochs	20
Validation Split	10%
Early Stopping	Patience = 5
Learning Rate Scheduler	ReduceLROnPlateau (factor = 0.5, patience = 3)

TABLE 2: CNN architecture and training configuration

ReLU, Rectified Linear Unit; CNN, Convolutional Neural Network

Results And Discussion

Experimental evaluation

The proposed CNN, along with three baseline machine learning classifiers, was tested using the official 10-fold cross-validation method from the UrbanSound8K dataset. All models were trained and tested on the same set of predefined folds to ensure a fair and consistent way of evaluating them. The traditional machine learning models used a carefully made set of features including MFCCs, Chroma, and Spectral features, while the CNN used log-Mel spectrograms directly as input.

The models were assessed using four different measures: accuracy, precision, recall, and macro F1-score. Accuracy shows how often the model correctly assigns samples to their right class. Precision measures how well the model identifies positive cases, and recall shows how well it finds all relevant instances. The macro F1-score, which is the average of the F1-scores for each class without weighting, was chosen as the main performance measurement because it treats each sound class equally, even when there are unequal numbers of samples in each class.

To check the reliability of the proposed approach, the average performance across all 10 folds of cross-validation was reported along with the standard deviation and 95% confidence interval. Also, paired t-tests and Wilcoxon signed-rank tests were done to see if the differences in performance between the CNN and the traditional machine learning models were statistically significant.

The results of the experiments are covered in the next sections. First, the performance of all the tested classifiers is compared using the chosen evaluation measures. Then there is a statistical analysis of these results, followed by a detailed discussion of how well the CNN performed, using a confusion matrix.

Performance comparison of classification models

The proposed CNN was tested against three traditional machine learning methods: KNN, Linear SVM, and RF. All models were assessed using the official 10-fold cross-validation method from UrbanSound8K, and the average results from all 10 folds are shown in Table *3*.

The CNN performed the best among all the models, achieving an accuracy of 69.35%, precision of 74.96%, recall of 71.22%, and a macro F1-score of 70.42%. These results show that learning hierarchical features directly from log-Mel spectrograms helps the CNN better understand the time and frequency patterns in sounds, which are not fully captured by the manually designed acoustic features used in this study.

Among the traditional models, RF gave the best results with an accuracy of 64.56%, precision of 67.63%, recall of 66.27%, and a macro F1-score of 65.72%. This model works by combining predictions from many decision trees, which makes it good at handling complex patterns while also reducing overfitting through techniques like bootstrap aggregating and random feature selection.

The Linear SVM had moderate performance, with an accuracy of 60.90%, precision of 62.79%, recall of 63.32%, and a macro F1-score of 62.04%. While it is efficient for high-dimensional data, it can't model complex, nonlinear relationships in environmental sounds as effectively as methods like RF or deep learning.

KNN had the worst performance, with an accuracy of 54.75%, precision of 59.49%, recall of 57.81%, and a macro F1-score of 56.26%. As an instance-based method, KNN relies heavily on how well the manually created features separate different sound types, and it can struggle when similar sounds have overlapping features.

Overall, the results show a progressive increase in classification performance from KNN to Linear SVM, RF, and finally the CNN. Relative to RF, the strongest classical baseline, the CNN improved classification accuracy by approximately 4.8 percentage points and macro F1-score by approximately 4.7 percentage points, indicating the benefit of automatically learned representations in this evaluation.

Model	Accuracy	Precision	Recall	Macro F1-score
K-Nearest Neighbours (KNN)	0.5475	0.5949	0.5781	0.5626
Linear SVM	0.609	0.6279	0.6332	0.6204
Random Forest	0.6456	0.6763	0.6612	0.6572
Proposed CNN	0.6935	0.7496	0.7122	0.7042

TABLE 3: Performance comparison of the evaluated models under the official UrbanSound8K 10-fold cross-validation protocol

SVM, Support Vector Machine; CNN, Convolutional Neural Network

Statistical analysis

To check if the improvements seen with the proposed CNN are statistically significant, paired t-tests and Wilcoxon signed-rank tests were used on the classification accuracies from the 10 cross-validation folds. The results are shown in Table *4*.

The t-test results show that the CNN performed significantly better than all three baseline classifiers. The biggest difference was compared to KNN, with a t-statistic of 6.3882 and a p-value of 0.000127. The next biggest difference was against the Linear SVM, with a t-statistic of 5.8422 and a p-value of 0.000246. Even though the difference with RF was smaller, it was still statistically significant, with a t-statistic of 3.6952 and a p-value of 0.004957. Since all p-values are below 0.05, we can reject the idea that the CNN performs the same as the other models.

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

The Wilcoxon signed-rank test, which does not require the performance differences to be normally distributed, also supports these findings. For all three comparisons, the Wilcoxon statistic was 0.0 with a p-value of 0.001953, showing the CNN performed consistently better across the cross-validation folds.

To better understand how consistent the model is, we calculated the mean performance, standard deviation, and 95% confidence intervals for each evaluation metric. These are listed in Table *5*. They give an idea of how much the model's performance might vary when tested on new, unseen data.

Model Comparison	Paired t-Statistic	p-value	Wilcoxon Statistic	p-value
CNN vs. Linear SVM	5.8422	0.0002	0	0.002
CNN vs. KNN	6.3881	0.0001	0	0.002
CNN vs. Random Forest	3.6952	0.005	0	0.002

TABLE 4: Statistical comparison of the CNN with the baseline machine learning models

KNN, K-Nearest Neighbours; SVM, Support Vector Machine; CNN, Convolutional Neural Network

Metric	Mean	Standard Deviation	95% Confidence Interval
Accuracy	0.6935	0.0781	0.6376 – 0.7494
Precision	0.7496	0.0761	0.6952 – 0.8041
Recall	0.7122	0.0752	0.6584 – 0.7659
Macro F1-score	0.7042	0.082	0.6455 – 0.7629

TABLE 5: Summary statistics of the proposed CNN over the 10-fold cross-validation

CNN, Convolutional Neural Network

The small standard deviations in all the evaluation measures show that the proposed CNN performed consistently across the 10 cross-validation folds. Also, the tight confidence intervals suggest that the performance estimates are stable, and the improvements seen are probably not due to random changes between the training and testing data. When combined with the results from the hypothesis tests, these results show that the proposed CNN offers better classification accuracy and dependable generalisation for environmental sound classification.

Confusion matrix analysis

Figure *1* presents the confusion matrix of the proposed CNN, providing class-level information about its predictions beyond the aggregate performance metrics reported in the previous section. Although the model attained an overall accuracy of 69.35%, Figure *1* shows variation in recognition performance across the 10 environmental sound classes, reflecting differences in their acoustic characteristics.

The proposed CNN achieved high recognition rates for several sound categories, as reflected by the dominant diagonal values in the confusion matrix. Dog bark recorded the highest number of correctly classified samples (826), followed by street music (762), children playing (758), drilling (721), siren (670), engine idling (614), jackhammer (592), air conditioner (433), gunshot (350), and car horn (330).

These results suggest that the CNN learned useful spectral-temporal representations for distinguishing most of the evaluated environmental sound classes despite the diversity of the acoustic recordings.

The confusion matrix also highlights several notable misclassification patterns. The largest source of confusion occurred between air conditioner and engine idling, where 219 air conditioner recordings were classified as engine idling and 104 engine idling recordings were classified as air conditioner. This behaviour is expected because both sound classes are characterised by continuous, low-frequency, and relatively stationary acoustic components, making them difficult to distinguish using spectral information alone.

Another significant confusion was observed between jackhammer and drilling. Specifically, 151 jackhammer samples were predicted as drilling, while 122 drilling samples were classified as jackhammer. Since both sound events originate from construction equipment and exhibit repetitive impulsive patterns with similar frequency characteristics, distinguishing between them remains challenging even for deep learning models. Likewise, 129 engine idling recordings were misclassified as jackhammer, suggesting that certain recordings shared overlapping low-frequency energy distributions.

The confusion matrix further indicates that 104 street music recordings were classified as children playing, while 59 dog bark recordings were also predicted as children playing. These errors are likely attributable to the presence of background urban noise and the complex acoustic mixtures commonly encountered in outdoor recordings, where multiple sound sources may coexist within a single audio clip.

Overall, the confusion matrix shows that the proposed CNN provides reliable classification across most of the evaluated environmental sound categories. Although confusion persists among acoustically similar classes, most predictions are concentrated along the principal diagonal, indicating effective class discrimination and good generalisation capability. These observations are consistent with the quantitative performance reported in Tables *3*-*5*, further confirming that the proposed CNN effectively captures discriminative acoustic patterns for environmental sound classification.

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

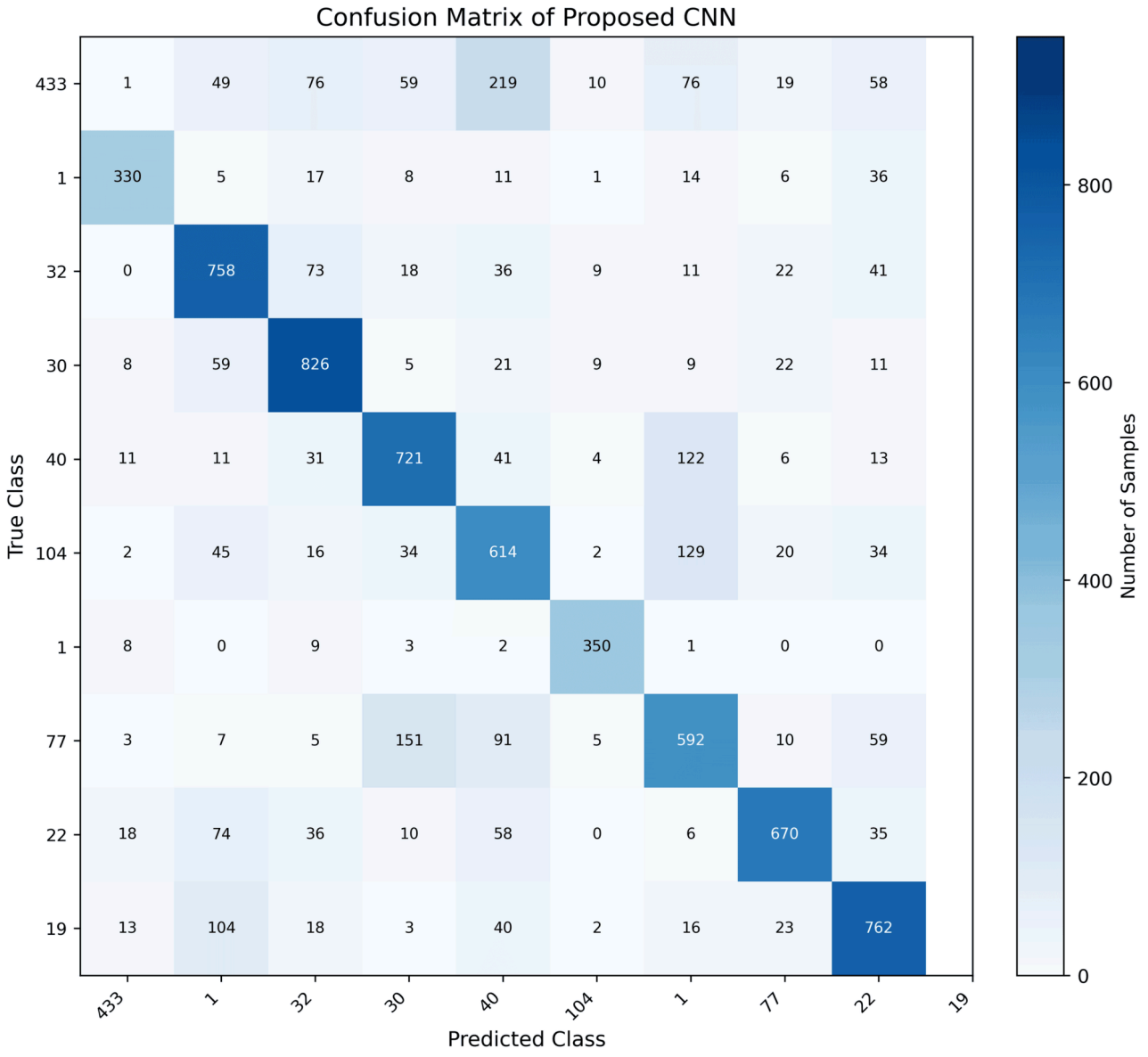

FIGURE 1: Confusion matrix obtained for the CNN under the official UrbanSound8K 10-fold cross-validation protocol

CNN, Convolutional Neural Network

Discussion

The experimental results show that the proposed CNN performed better than the three baseline machine learning classifiers in all the evaluated metrics. Table *3* shows that the CNN had the highest accuracy (69.35%) and macro F1-score (70.42%). The statistical analysis section also shows that these differences are statistically significant. These results suggest that the deep learning approach used here was more effective than the traditional handcrafted-feature methods tested in this study.

The observed advantage of the CNN is in line with recent research in ESC that has been focusing more on learned spectro-temporal representations, attention-based time-frequency modeling, metric and contrastive learning, and modern CNN and transformer architectures, rather than only using handcrafted features [4-6]. This learned representation enables the model to capture detailed temporal and spectral patterns that may not be fully represented by the handcrafted features used in the baseline models, which leads to better classification performance across various types of environmental sounds.

Recent studies in journals also show a trend towards richer spectro-temporal modeling techniques, including interactive time-frequency attention, metric and contrastive learning, advanced CNN and transformer architectures, and feature-fusion strategies for environmental sound classification [4-7].

Even though the proposed CNN showed better overall performance, the confusion matrix analysis showed that some acoustically similar sound categories are still hard to distinguish. There was notable confusion between air conditioner and engine idling, as well as between jackhammer and drilling. These sounds share similar frequency patterns and temporal structures, making them naturally difficult to separate, especially in recordings with background noise or overlapping sounds. These findings suggest the need to explore recent ESC strategies that use attention-based time-frequency modeling, modern CNN and transformer architectures, feature-fusion mechanisms, metric and contrastive learning, and data augmentation to improve the ability to distinguish between acoustically similar classes [4-7].

Even though there were some limits, the model still worked well in all 10 cross-validation tests. This is shown by the low standard deviations and tight 95% confidence intervals in Table *5*. Also, statistical tests proved that the improvements were real and not just because of chance in how the data was split. These results show that the CNN is reliable and can work well in different situations for environmental sound classification.

The framework gives a good balance between how well it classifies sounds and how complex the model is. It uses standard log-Mel spectrogram inputs with a simple convolutional structure. Because of this, the CNN is a useful method for classifying environmental sounds and can be used in areas like city noise monitoring, security systems, and audio analysis that responds to the environment.

To promote transparency and reproducibility, the source code and supplementary materials used in this study have been made publicly available in an online repository [10] (Table *6*).

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

Study	Dataset	Method	Key Difference from Present Study
Salamon et al. [1]	UrbanSound8K	Benchmark dataset	Introduced the UrbanSound8K benchmark dataset used in this study.
Piczak [2]	UrbanSound8K	CNN	Demonstrated CNN-based environmental sound classification using spectrogram representations.
Hershey et al. [3]	AudioSet	Deep CNN	Focused on large-scale audio representation learning using the AudioSet dataset.
Chen and Peng [4]	ESC-50, UrbanSound8K	Two-stream DNN + interactive time-frequency attention	Uses separate time- and frequency-oriented branches with inter-branch interaction and depthwise separable convolution to model time-frequency dependencies while reducing computational complexity.
Yurtsever [6]	UrbanSound8K	Modern CNN architectures + ViT + augmentation	Evaluates modern CNN and Vision Transformer architectures with MixUp and SpecAugment under the official 10-fold UrbanSound8K protocol.
Zeng, Zhou and Cai [7]	UrbanSound8K, ESC-50	Feature fusion + SResNet	Combines complementary linear and nonlinear feature representations and investigates their integration with an attention-enhanced SResNet framework.
Bansal and Garg [11]	Environmental sound datasets	CRNN	Proposed a CRNN architecture for environmental sound classification.
Mushtaq and Su [12]	Environmental sound datasets	Regularized CNN	Investigated CNN robustness using regularization and data augmentation.
Tokozume and Harada [13]	Environmental sound datasets	End-to-End CNN	Learned discriminative representations directly from audio using an end-to-end CNN.
Present Study	UrbanSound8K	Classical ML + CNN	Provides a systematic comparison of classical machine learning models and CNNs using identical preprocessing, official 10-fold cross-validation, statistical

			significance testing, and confidence-interval analysis.

TABLE 6: Comparison of the present study with representative environmental sound classification studies

CNN, Convolutional Neural Network; ML, Machine Learning; DNN, Deep Neural Network; ESC, Environmental Sound Classification; ViT, Vision Transformer; SResNet, Squeeze-and-Excitation Residual Network; CRNN, Convolutional Recurrent Neural Network

Limitations

Although the proposed CNN achieved superior performance compared with the evaluated classical machine learning models, several limitations remain. First, the experimental evaluation was conducted exclusively on the UrbanSound8K dataset; therefore, the generalisability of the proposed approach to other environmental sound datasets requires further investigation. Second, the study considered a conventional CNN architecture and did not evaluate newer approaches such as transformer-based architectures, attention-enhanced networks, metric and contrastive learning, or feature-fusion methods [4-7]. Third, although the GUI demonstrates the practical deployment of the proposed system, it was not evaluated through user studies or real-time performance benchmarking. Finally, no data augmentation techniques were employed during training, and incorporating advanced augmentation strategies may further improve robustness for acoustically similar sound classes.

Conclusions

This study looked at how well classical machine learning methods and a CNN work for classifying environmental sounds using the UrbanSound8K dataset. They tested KNN, Linear SVM, and RF, along with a CNN that was trained using log-Mel spectrograms. All models were tested using the official 10-fold cross-validation method. The CNN performed best, with an accuracy of 69.35%, precision of 74.96%, recall of 71.22%, and a macro F1-score of 70.42%, which was better than the results from the classical models. Statistical tests showed that the performance differences were consistent across all the cross-validation folds, not just due to chance.

Looking at the confusion matrix, most environmental sound categories were classified correctly, but there was confusion between similar-sounding classes like air conditioner and engine idling, or jackhammer and drilling. These results show that distinguishing between similar sounds in complex urban environments is still a challenge and there's room for improvement in how well these sounds can be told apart.

The findings suggest that the CNN's ability to learn from the spectral and temporal patterns in the sound data gave it better classification results than the classical models, which relied on manually created features. However, RF still holds its own when being efficient with computation and easy to understand is important.

Looking ahead, future work will explore recent methods in environmental sound classification, such as attention-based models and transformer structures, along with techniques like metric learning, contrastive learning, feature fusion, and data enhancement. A special focus will be on improving the ability to tell apart sounds that are acoustically similar. Also, testing these models on bigger and more varied datasets, and making them work efficiently on devices with limited resources, could be important directions for future research.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Sushant Sharma, Rushali R. Katkar, Sushama A. Shirke, Shreya Jare

Acquisition, analysis, or interpretation of data: Sushant Sharma, Rushali R. Katkar, Sushama A. Shirke, Shreya Jare

Drafting of the manuscript: Sushant Sharma, Rushali R. Katkar, Sushama A. Shirke, Shreya Jare

Critical review of the manuscript for important intellectual content: Sushant Sharma, Rushali R. Katkar, Sushama A. Shirke, Shreya Jare

Supervision: Rushali R. Katkar, Sushama A. Shirke

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in Google Drive at https://drive.google.com/drive/folders/1cunG1susJ40dH6lSa30naW7UvOvjxjLi?usp=drive_link, reference NA.

References

1. Salamon J, Jacoby C, Bello JP: A dataset and taxonomy for urban sound research. Proceedings of the ACM International Conference on Multimedia. 2014, 1041-1044. 10.1145/2647868.2655045
2. Piczak KJ: Environmental sound classification with convolutional neural networks. Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP). 2015, 1-6. 10.1109/MLSP.2015.7324337
3. Hershey S, Chaudhuri S, Ellis DPW, et al.: CNN architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017, 131-135. 10.1109/ICASSP.2017.7952132
4. Chen B, Peng J: Environmental sound classification using two-stream deep neural network with interactive time-frequency attention. Applied Acoustics. 2025, 238:110794. 10.1016/j.apacoust.2025.110794
5. Chen F, Zhu Z, Sun C, Xia L: Evaluating metric and contrastive learning in pretrained models for environmental sound classification. Applied Acoustics. 2025, 232:110593. 10.1016/j.apacoust.2025.110593
6. Yurtsever U: Modern deep learning architectures for urban sound classification on UrbanSound8K. Sakarya University Journal of Computer and Information Sciences. 2026, 9:934-942. 10.35377/saucis...1830726
7. Zeng J, Zhou C, Cai D: A deep learning framework for environmental sound classification by fusing linear and nonlinear features. Journal on Advances in Signal Processing. 2026, 2026:37. 10.1186/s13634-026-01299-y
8. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O: librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference (SciPy). 2015, 18-25. 10.25080/Majora-7b98e3ed-003
9. Pedregosa F, Varoquaux G, Gramfort A, et al.: Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011, 12:2825-2830.
10. Sharma S, Katkar R, Shirke SA, Jare S: Source code and dataset for: Analysis of Classical and Deep Learning Models for Environmental Sound Classification. Google Drive. 2026,
11. Bansal A, Garg NK: Robust technique for environmental sound classification using convolutional recurrent neural network. Multimedia Tools and Applications. 2024, 83:54755-54772. 10.1007/s11042-023-17066-2
12. Mushtaq Z, Su SF: Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Applied Acoustics. 2020, 167:107389. 10.1016/j.apacoust.2020.107389
13. Tokozume Y, Harada T: Learning Environmental Sounds with End-to-End Convolutional Neural Network. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017, 2721-2725. 10.1109/ICASSP.2017.7952651

How to cite this article:
Sharma S, Katkar R R, Shirke S A, et al. (September 16, 2026) Environmental Sound Classification for Audio-Based Action Recognition: A Comparative Study of Classical and Deep Models. Cureus J Comput Sci 3 : es44389-026-00281-x. DOI https://doi.org/10.7759/s44389-026-00281-x