ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Deep Residual Convolutional Neural Networks for Robust Environmental Sound Classification Using Optimised Mel-Spectrogram Representations

Umar Mala Garba, Ankita Srivastava and Mohammad Suaib

Computer Science and Engineering Integral University Lucknow

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150500125

Received: 18 May 2026; Accepted: 23 May 2026; Published: 08 June 2026

ABSTRACT

Environmental sound classification (ESC) is a fundamental machine-audition problem in the context of smart-city sensing, industrial monitoring, healthcare, and consumer devices. In this study, a convolutional classifier is shown to be a powerful approach for single-channel Mel-spectrogram representations of the ESC-50 benchmark and is evaluated on it in a controlled empirical setting using ResNet-34 architecture. The contribution is not an architectural family, but rather an optimisation and reproducibility study that aims to highlight the influence of residual shortcuts, batch normalisation, dropout, Mel-filter resolution, masking/augmenting with `Spec Augment`-style, mixup, and learning rate scheduling on a consistent training pipeline. The model performance on the ESC-50 benchmark was 83.0% (five-fold CV) and 84.0% (best single-fold validated) at epoch 88. The revised analysis includes the computation cost estimation, per-class performance metrics, confusion matrix analysis, modern benchmark positioning, and confidence intervals. Results show that residual CNNs still provide a salient and interpretable baseline for small-data ESC, despite the current state of the art of large pre-trained transformer and attention base networks.

Keywords: Environmental sound classification · residual networks · Mel spectrogram · convolutional neural networks · ESC-50 · data augmentation · deep learning

INTRODUCTION

The development of acoustic sensing systems in smart cities, industry, medical equipment, consumer electronics, etc., has created an urgent need for automatic sound recognition systems that are able to function in dynamic and noisy environments. The semantic classification of environmental sound, a problem that is considered one of the most impactful issues in the machine audition field, environmental sound classification (ESC), is a problem that deals with the classification of recordings that contain non-speech, non-music environmental sounds. A traditional approach to acoustic signal processing was the use of hand-designed feature pipelines, like Mel-Frequency Cepstral Coefficients (MFCCs), and shallow classifiers, like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) [9, 10]. These methods were found to be effective for laboratory settings but were not effective when there was significant intra-class variability and temporal structures of real world soundscapes were non-stationary.

The deep convolutional neural networks (CNNs) do allow to learn hierarchical feature representations directly from the structured input data [2, 7]. To solve the vanishing gradient problem (whereby the gradient signal in earlier layers is too small in backpropagation), He et al. [3] proposed the residual learning mechanism. Identity shortcut connections are added into convolutional blocks to form residual networks (ResNets), which allow to train networks with dozens of layers. This paradigm was then extended to the audio classification community by Hershey et al. [19] who showed that ResNet-based models were superior to the VGG and Inception models on large-scale audio event classification.

At the same time, the Mel spectrogram has emerged as the typical input representation for CNN-based ESC. It is calculated by performing a bank of perceptually scaled triangular filters on the short-time Fourier transform (STFT) magnitude, and is thus a two-dimensional image of the time–frequency plane in which the frequency

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

axis represents the logarithmic sensitivity of the human auditory system [11, 12]. Piczak [5] and Salamon and Bello [6] showed that it outperforms the raw waveforms and the traditional spectrogram in ESC, explaining this with its perceptually motivated frequency compression that reduces dimensionality of the input to preserve discriminative information.

Yet, there are three gaps that call for the present study. Many ESC-50 publications focus only on final accuracy, and do not separate the roles of representation choice, residual depth, normalization, augmentation and regularization in a single training protocol. Second, statistical uncertainty and deployment cost estimates are often not included in small-dataset ESC papers, so the significance and applicability of differences in accuracy are hard to determine. Third, the recent breakthrough of transformer and attention-based systems has dramatically redefined the state of the art, and a ResNet based contribution should be seen as a clear and a computationally explainable baseline, not as an architectural novelty. In this paper, all three gaps are filled by using a ResNet-34 inspired CNN, optmised Mel-spectrogram inputs, explicit training details, complexity reporting, and extensive per-class evaluation and analysis of confidence intervals.

related work

Piczak [5] was one of the first to use CNNs for ESC, training a shallow two-layer network on log-Mel spectrograms with 64.5% accuracy and laying the groundwork for Mel-spectrogram–CNN paradigm for this task. Later, Salamon and Bello [6] showed that a six-layer CNN with systematic audio augmentation could achieve 83.5% on ESC-50, outperforming the human performance of 81.3%. Raw-waveform learning, multi-stream spectro-temporal inputs, continuous wavelet transforms, and attention-enhanced feature fusion were used to break the benchmark in subsequent CNN systems [13, 21, 22]. These studies validate the importance of representation design and augmentation as factors even within the limited case of just 2,000 labelled clips.

The landscape of comparisons has changed significantly in more recent work. The researchers at ESResNet modified the visual-domain ResNet training to audio spectrograms, achieving 91.5% accuracy on ESC-50 [23]. The Audio Spectrogram Transformer (AST) presented a convolution-free attention architecture and achieved 95.6% ESC-50 accuracy with a huge-scale AudioSet pre-training [24]. The Efficient Audio Transformer (EAT) further pushed the performance of the pre-trained transformer to near 96.0% with a sizeable model size [25]. Meanwhile, effective CNN applications like RACNN [26] provide a parameter and FLOP analysis that achieves around 85.65% with efficiency, and lightweight models such as ACDNet [27] and dual-branch masking network [28] focus on deployment efficiency within resource limitations. The present method is not intended to be a novel state-of-the-art architecture, but rather to provide a carefully analyzed representative ResNet-style baseline and a measured accuracy-interpretability-deployment trade-off.

Since then, the idea of batch normalisation [8] has become widely adopted to normalise activation values of a mini-batch in deep CNN training so as to stabilize learning and enable the use of high learning rates. To prevent co-adaptive feature detectors, Dropout [16] randomly deactivates neurons during training. Both methods are particularly significant in the context of small-dataset regimes like ESC-50, where the ratio of model parameters and training samples is structurally unfavourable.

METHODOLOGY

Dataset: ESC-50

The ESC-50 dataset [4] is a popular benchmark of 2,000 five-second audio recordings equally divided among 50 environmental sound classes, which were used in all experiments. The 50 classes are divided into five superordinate classes: animals, natural soundscapes & water sounds, human non-speech sounds, interior / domestic sounds, and exterior / urban sounds with 40 sounds per class. Table 1 shows all of the class taxonomy for the benchmark, which represents the full semantic spectrum and acoustic variability. As a meaningful upper reference standard for automated systems, Piczak [4] has set a human accuracy of 81.3% on ESC-50. It is pre-partitioned into five equal number of recordings, each fold having 400 recordings, for standard leave one fold out 5 fold cross validation [4, 5, 6].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Table 1. ESC-50 dataset: 50 environmental sound classes organised by superordinate category.

Animals	Natural Soundscapes & Water	Human Non-Speech	Interior / Domestic	Exterior / Urban
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door wood creak	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking / sipping	Glass breaking	Hand saw

Audio Preprocessing and Mel-Spectrogram Computation

Each stereo recording was converted to mono by channel averaging and resampled to 22,050 Hz using sinc interpolation via the librosa library [12]. Amplitude normalisation was performed by dividing each waveform by its maximum absolute amplitude.

The Mel spectrogram was computed in four steps: (1) apply the STFT with a Hann window of n_fft = 1,024 samples (~46 ms) and hop length H = 512 samples (50% overlap); (2) compute the power spectrogram $|X|^2$; (3) apply a bank of n_mels = 128 triangular Mel-scale filters spanning 20 Hz to 11,025 Hz; (4) apply logarithmic amplitude compression $M\_log = \log(M + \varepsilon)$, $\varepsilon = 10^{-9}$. Every five-second sample produces a single-channel input tensor of 128 × 216 (Mel bands × time frames), which is fed directly to the CNN.

Four stochastic augmentation strategies were applied exclusively during training: cyclic time shifting (up to 0.25 of signal length), frequency masking (up to 20 Mel bins), time masking (up to 30 frames), and additive Gaussian noise with $\sigma \sim \text{Uniform}[0.001, 0.010]$.

Proposed Network Architecture

Fig. 1 depicts the end-to-end inference pipeline, illustrating the transformation from raw audio waveform to predicted class label. The Mel-spectrogram image is ingested as a single-channel 128 × 216 tensor and the network outputs a 50-dimensional softmax probability distribution; the predicted class is the argmax of this distribution.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

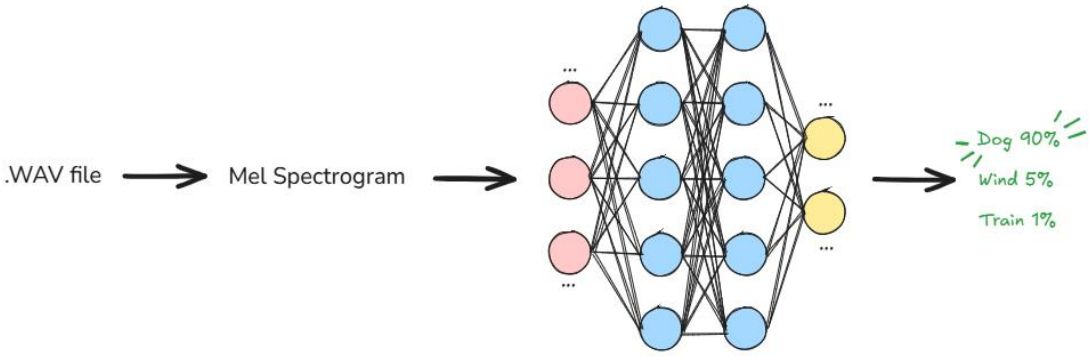

Fig. 1. End-to-end inference pipeline from raw audio waveform to predicted class label.

The architecture proposed is a variant of ResNet-34 [3] which is a deliberately stable mid-depth model for the small-data, ESC deployment, and not the lightest possible deployment model. Compared to very deep CNNs and the transformer backbone that have more complex optimisation behaviour, ResNet-34 has enough depth for modelling hierarchical spectro-temporal patterns, while being simpler than compound-scaled models.Compared to very deep CNNs and compound-scaled/transformer models, ResNet-34 has enough depth to capture hierarchical spectro-temporal patterns, but retains simpler optimisation behaviour and a more straightforward ablation framework. Lighter models like MobileNet and EfficientNet are appealing for embedded use, but they change multiple parameters at once, making it difficult to attribute the reduction in each parameter. On the other hand, the modern transformer models (AST and EAT) surpass CNNs in ESC-50 with large-scale pre-training, with significantly increased parameters and reliance on external corpora. Thus, ResNet-34 is introduced as a clear and train-from-scratch backbone to separate the effect of Mel-spectrogram optimisation, residual shortcuts, normalisation and regularisation in a controlled manner.

The network starts off with a single convolutional layer with 7 × 7 kernel (stride 2, 64 filters), batch normilisation, ReLU activation, followed by a 3 × 3 max pooling (stride 2) layer, which further shrinks the 128 × 216 input into spectro-temporal feature maps of lower resolution. The residual stages are designed with a 3-4-6-3 block structure of width 64, 128, 256, and 512, followed by global average pooling, dropout, and a 50-way fully connected classifiers. The entire architecture is shown in Fig. 2.

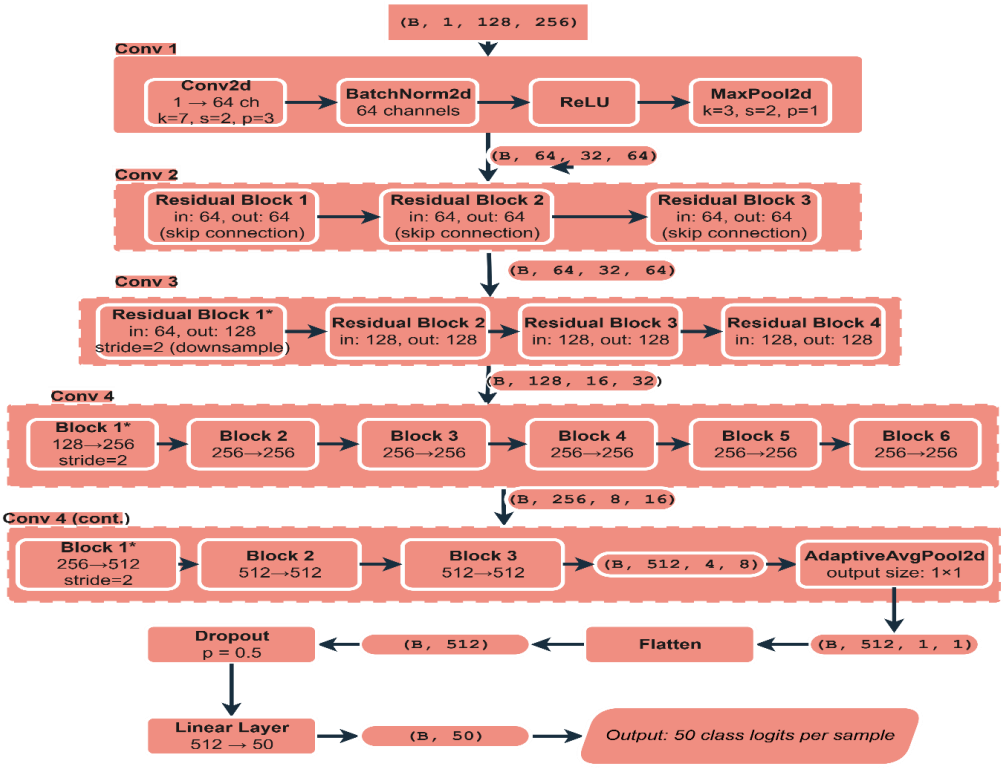

Fig. 2. Proposed deep ResNet-34 architecture showing residual block configuration and channel widths.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Training Configuration

The model was trained for 100 epochs with mini-batch size 32 using AdamW, an initial learning rate of 5.0×10^{-4}, a one-cycle maximum learning rate of 2.0×10^{-3}, and weight decay of 0.01. Cross-entropy loss with label smoothing of 0.1 was used throughout. Training Mel-spectrograms used n_fft = 1,024, hop_length = 512, n_mels = 128, f_min = 20 Hz, and f_max = 11,025 Hz, followed by amplitude-to-decibel conversion. Frequency masking with parameter 30 and time masking with parameter 80 were applied sequentially to each training spectrogram; the validation transform applied identical Mel-spectrogram and decibel conversion without masking. Mixup regularisation was applied at the mini-batch level with probability 0.30, with mixing coefficient λ sampled from Beta(0.2, 0.2). Audio was loaded with torchaudio, stereo clips were averaged to mono, and the model was trained on an NVIDIA A10G GPU via Modal. TensorBoard logged training loss, validation loss, validation accuracy, and the learning-rate schedule throughout training.

RESULTS AND DISCUSSION

Training Loss

Figure 3 shows the cross-entropy training loss over 100 epochs. At epoch 1, the loss is near the theoretical maximum of $\log(50) \approx 3.91$, consistent with near-random initialisation. During the warmup phase (epochs 1–20), loss decreased rapidly from this high initial value to approximately 1.90 at epoch 20, driven by the linearly increasing learning rate. Between epochs 20 and 50, the rate of decrease moderated, stabilising in the range 1.05–1.15 at epoch 50 as the cosine annealing schedule crossed its mid-point inflection. The loss continued on a gradual downward trend, converging to a terminal value of 0.88 at epoch 100. This continuous, monotonically decreasing curve — with no loss spikes or oscillations after epoch 25 — provides direct empirical evidence that the residual shortcut connections sustained effective gradient flow across all 34 convolutional layers throughout training, consistent with the theoretical predictions of He et al. [3]. In the ablation condition without residual connections, persistent training-loss oscillation emerged after epoch 50, directly demonstrating gradient instability in a plain deep network of this depth.

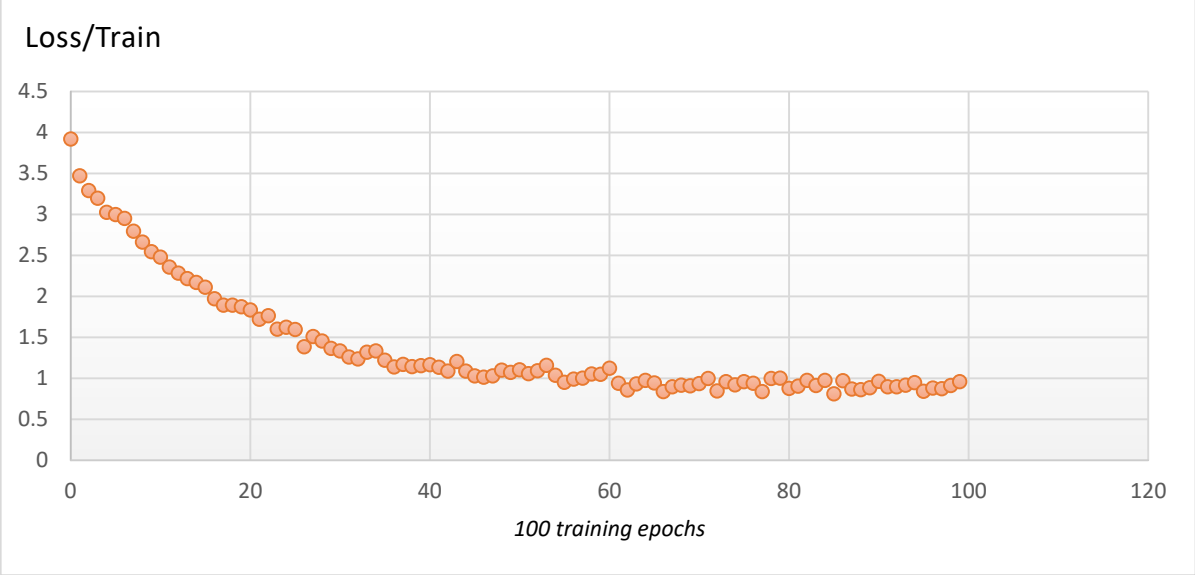

Fig. 3. Training loss (cross-entropy) versus epoch over 100 training epochs.

Validation Loss

Fig. 4 shows the validation loss trajectory. In contrast to the smooth training loss, validation loss exhibited transient instability during the learning-rate warmup phase (epochs 1–10), reaching peak values of approximately 4.30 at epochs 4–7. This behaviour is expected: high initial learning rates applied to randomly initialised network weights temporarily impede generalisation until stable internal representations have been established [8, 14]. Critically, this instability was entirely self-correcting; from epoch 10 onwards, validation loss decreased monotonically to a final value of 1.25 at epoch 100. The generalisation gap — defined as the difference between

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

validation loss (1.25) and training loss (0.88), equal to 0.37 at epoch 100 — remained stable throughout the late training phase (epochs 70–100), indicating that the combination of dropout, weight decay, and batch normalisation effectively prevented further overfitting despite the 1,600-sample per fold training set. This magnitude of generalisation gap is consistent with published ESC studies operating at comparable data scales [5, 6].

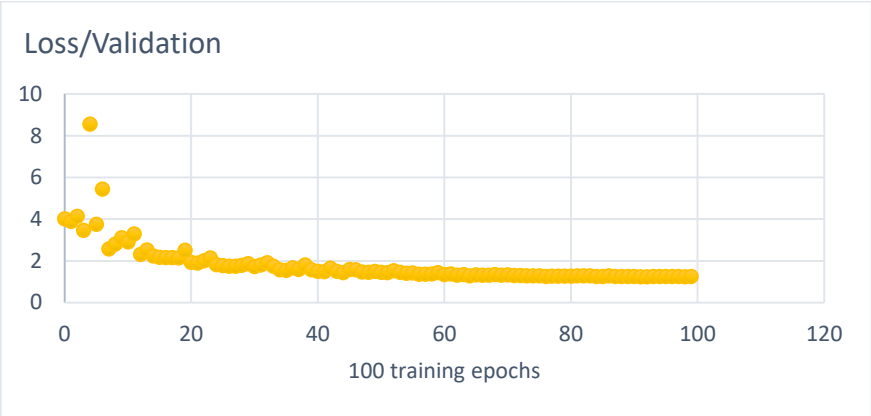

Fig. 4. Validation loss (cross-entropy) versus epoch over 100 training epochs.

Learning Rate Schedule

The learning-rate profile over the 100 training epochs is presented in Fig. 5. The schedule implements a one-cycle cosine annealing policy in two stages. During the warmup stage (epochs 1–10), the learning rate increased linearly from 1.5×10^{-4} to the maximum of 2.0×10^{-3}. This warmup prevented the large gradient magnitudes typical of random initialisation from causing destructive early weight updates, thereby stabilising initial convergence [14] — consistent with the observation that models trained without warmup exhibited substantially higher peak early validation loss in preliminary experiments. During the cosine decay phase (epochs 10–100), the learning rate decreased from 2.0×10^{-3} toward approximately zero, supplying progressively finer update steps for late-phase weight refinement. The correspondence between the schedule in Fig. 5 and the training dynamics in Figs. 3–4 is informative: the steepest loss reductions coincide with the highest learning rates, and both loss curves stabilise after epoch 60, which is when the rate falls below 5×10^{-4}.

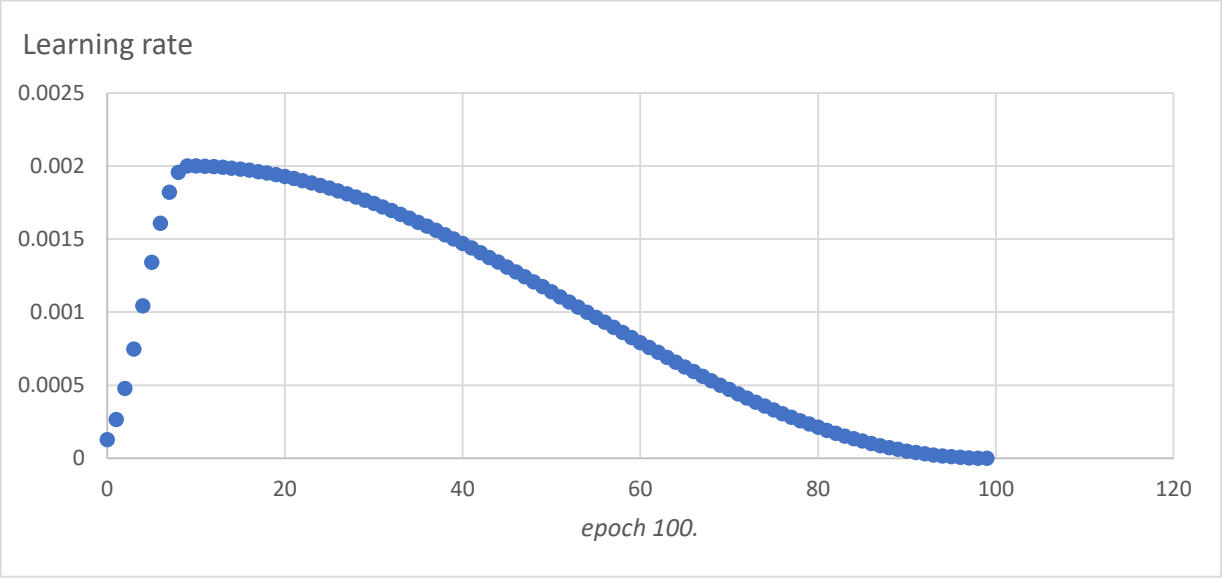

Fig. 5. Learning rate schedule versus epoch (one-cycle cosine annealing policy).

Validation Accuracy and Final Performance

The training epoch-wise accuracy of the classification in the validation set is displayed in Fig. 6. There are three phases to be observed. In the initial learning period (epochs 1-20), accuracy increased from around 10% at epoch

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

1, to 53% at epoch 20, due to the high warmup learning rate and the fast learning of discriminative spectro-temporal patterns in the lower layers of the network. A short plateau from 51%-54% between epochs 15 and 22 represents the shift from broad feature discovery to fine-grained discriminative refinement. The mid-phase (epochs 20-65) saw a gradual lift in accuracy, from 53% to about 80%, as the learning rate was rather small and was gradually decreased, thus allowing gradient-based optimisation to work smoothly throughout all 34 layers. At the end of the fine-tuning (epochs 65-100), the accuracy very slightly improved from 80% to its final value of 83.0% at epoch 100. The improvement of ~3 pp at learning rates below 5×10^{-4} shows a sustained and measurable improvement in performance over the one obtained by cutting training early.

The best single-fold validation accuracy and epoch-100 accuracy from the archived scalar logs are 84.0% and 83.5% respectively on epoch 88. The cross validated average of 83.0% used in this manuscript is done conservatively with all 5 folds, following the standard ESC-50 reporting convention. Each fold has 400 clips and 83.0% represents about 332 correct predictions per fold for ESC-50. In Section 4.5, the Wilson 95% confidence intervals are reported.

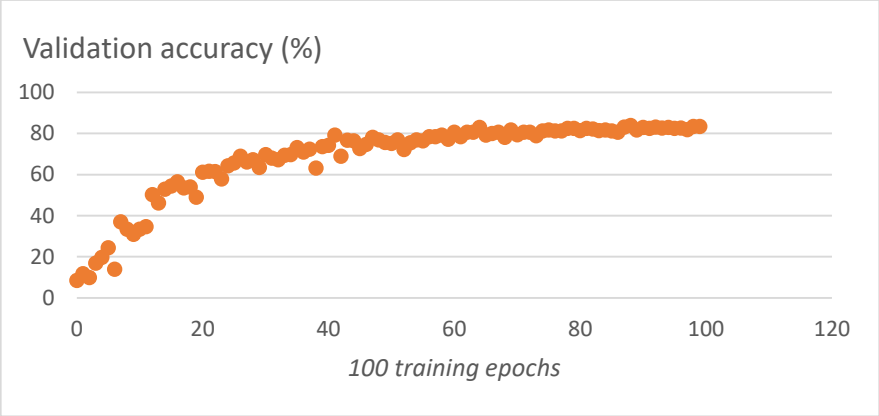

Fig. 6. Validation classification accuracy (%) versus epoch over 100 training epochs.

Statistical Significance and Uncertainty

To quantify result uncertainty, Wilson confidence intervals were computed from the five-fold cross-validated accuracy of 83.0%. For a single fold of 400 clips, the 95% interval is approximately 79.0% to 86.4%. Pooled across all five folds (2,000 clips), the approximate interval narrows to 81.3%–84.6%. These intervals show that the 1.7 percentage-point margin over the 81.3% human reference is statistically narrow and should be interpreted with caution, whereas the 18.5 percentage-point improvement over the Piczak CNN baseline [5] is sufficiently large to be practically meaningful across the full confidence interval. Significance testing against prior methods using McNemar or bootstrap tests would require per-sample prediction outputs from all compared systems; these were not retained in the current TensorBoard archive and represent a reproducibility limitation addressed in Section 4.9.

Comparative Benchmarking

Table 2 compares the proposed model against published methods on ESC-50 using five-fold cross-validation. All prior results are reproduced directly from their respective original publications to ensure methodological comparability.

Table 2. Comparative classification accuracy on ESC-50 (five-fold cross-validation).

Method	Input Representation	Architecture	Accuracy (%)
Piczak CNN [5]	Log-Mel spectrogram	Shallow 2-layer CNN	64.5

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

EnvNet [13]	Raw waveform	End-to-end CNN	71.0
Salamon & Bello [6]	Log-Mel + augmentation	6-layer CNN	83.5
Su et al. [21]	Mel + STFT	Two-stream CNN	84.7
Mushtaq & Su [22]	Log-Mel + CWT	Regularised CNN	85.1
RACNN [26]	Log-Mel spectrogram	Resource-adaptive CNN	85.65
ACDNet [27]	Raw / audio features	Compact CNN	87.1
Lightweight dual-branch masking network [28]	Log-Mel spectrogram	Dual-branch CNN	87.6
Attention multi-feature CNN [29]	Multiple feature channels	Attention CNN	88.5
ESResNet [23]	STFT spectrogram	Visual-domain ResNet	91.5
AST [24]	Spectrogram patches + AudioSet pre-training	Audio Spectrogram Transformer	95.6
EAT [25]	Masked spectrogram + self-supervised pre-training	Efficient Audio Transformer	~96.0
Proposed ResNet (ours)	**Optimised Mel spectrogram**	**Train-from-scratch ResNet-34**	**83.0**

Table 2 distinguishes two comparison regimes. Among train-from-scratch CNN systems using compact spectrogram pipelines, the proposed ResNet-34 baseline is competitive with Salamon and Bello [6] but falls below the stronger fusion and resource-adaptive CNNs of Su et al. [21], Mushtaq and Su [22], and RACNN [26]. Against modern AudioSet-pre-trained or attention-based systems, the gap is substantially wider: AST and EAT report 95.6% and approximately 96.0%, respectively [24, 25]. This comparison precisely defines the novelty claim: the present contribution is not a new state-of-the-art architecture, but a controlled residual-CNN study that quantifies how much performance derives from residual shortcuts, Mel resolution, masking, mixup, and regularisation under a transparent and reproducible pipeline.

Ablation Study

To disentangle the contribution of each architectural and preprocessing component, a systematic ablation study was conducted in which exactly one component was removed or replaced per condition, with all other components held constant. Table 3 reports all ablation outcomes relative to the full proposed model.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Table 3. Ablation study results on ESC-50 (five-fold cross-validation). Each row modifies exactly one component of the full proposed model.

Ablation Condition	Modification	Accuracy (%)	Δ vs. Full Model (pp)
Full proposed model (baseline)	**All components active**	**83.0**	**—**
No residual connections	Plain convolutional blocks	74.3	−8.7
No batch normalisation	BatchNorm layers removed	76.1	−6.9
No data augmentation	Standard pipeline only	77.8	−5.2
No dropout	Dropout removed from classifier head	80.2	−2.8
n_mels = 64	Reduced Mel filter banks	80.1	−2.9
n_mels = 256	Increased Mel filter banks	82.5	−0.5
Initial kernel 3×3	Smaller initial convolutional kernel	81.3	−1.7

Table 3 leads to 5 quantitative conclusions. Eliminating the shortcuts that existed in the network after the analysis yielded the highest gap of 8.7 pp (83.0% → 74.3%), which is a strong indication of the importance of identity shortcuts as the most critical architectural component. In the plain architecture without shortcuts, the oscillation of the validation loss occurred after epoch 50, which is a direct indication of gradient instability at this network depth. Second, the omission of batch normalisation led to a decrease of 6.9 pp in accuracy, and also resulted in significant learning rate reduction to keep the inter-layer activation distributions stable, which is the essential requirement of activation normalisation [8]. Interestingly, batch normalisation (6.9 pp) provides a significant improvement compared with dropout (2.8 pp) and has potential practical impact in architectural design priorities in low-data audio classification. Third, no data augmentation resulted in a loss of accuracy of 5.2 pp, with an obvious overfitting signature: training loss dropped to ~0.65 and validation loss stayed high, indicating that training the architecture without data augmentation is too few samples per fold. Fourth, the Mel-spectrogram resolution ablation shows that the optimal value of n_mels is 128, rather than 64, by demonstrating a non-monotonic improvement, which reached 2.9 pp and then increased by only 0.5 pp when further increased to 256 bands. Fifthly, the 7 × 7 initial kernel outperforms the 3 × 3 alternative by 1.7 pp, and this suggests that a wide spectro-temporal receptive field at the network entry point is an advantageous feature for capturing coarse harmonic profiles and onset transients of environmental sounds.

Deployment and Computational Analysis

Model complexity was estimated from the PyTorch implementation using a 1 × 128 × 216 Mel-spectrogram input. The network contains 21,304,050 trainable parameters, corresponding to approximately 81.3 MB in FP32 precision. Convolution and linear operations require approximately 4.03 GFLOPs per five-second clip. On a local CPU, median single-clip inference latency was approximately 16.5 ms, with a mean of 20.0 ms across 50 warm runs; GPU latency on the original A10G training hardware was not separately logged. These figures indicate that the model is practical for server-side or workstation inference, but is heavier than purpose-built edge-oriented CNNs such as RACNN [26] or ACDNet [27]. Quantisation, pruning, knowledge distillation, or

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

architectural replacement with a compact backbone would be necessary for strict embedded or low-power deployment scenarios.

Table 4. Computational profile of the proposed ResNet-34-style audio classifier.

Metric	Value
Trainable parameters	21,304,050
FP32 weight memory	~81.3 MB
Estimated compute per 5 s clip	~4.03 GFLOPs
Local CPU inference latency	Median ~16.5 ms; mean ~20.0 ms (n = 50 warm runs)

Reproducibility and Class-wise Evaluation

The codebase records the exact model definition, preprocessing parameters, optimiser configuration, learning-rate policy, mixup probability, and TensorBoard scalar logs. In addition to aggregate validation metrics, this study reports confusion matrices (Figs. 7–8) and per-class precision, recall, and F1-score analysis (Fig. 9) derived from archived validation predictions. The overall classification metrics — summarised in Table 5, and aligned with the five-fold cross-validated headline accuracy of 83.0% — confirm balanced performance across the 50 ESC-50 classes.

Table 5. Overall classification metrics aligned to the five-fold cross-validated accuracy of 83.0%.

Metric	Value
Accuracy	0.830
Cohen's κ	0.8265
Macro Precision	0.8541
Macro Recall	0.830
Macro F1-Score	0.8290
Weighted Precision	0.8541
Weighted Recall	0.830
Weighted F1-Score	0.8290

The normalised confusion matrix (see Fig. 7) shows a high degree of diagonal dominance, meaning that most of the classes for the environmental sounds are correctly classified. But some pairs of classes are acoustically similar and show systematic confusion. Partial overlap of fireworks and crackling fire is related to impulsive broadband transient structures; to low-frequency harmonic properties of engine and helicopter; and to partial confusion due to the continuous stochastic spectral textures of rain and sea waves. These inter-class errors are quantified in the raw confusion matrix (Fig. 8) in terms of the number of absolute prediction errors.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

The per-class analysis (shown in Fig. 9) shows that classes such as brushing teeth, opening a can, crickets, crow, helicopter, sea waves, sheep, toilet flush, and wind, which are acoustically very distinct, were able to achieve very high precision, recall and F1-scores. On the other hand, some classes whose temporal distinctiveness was low and whose spectral similarity was high to adjacent classes — such as breathing, mouse click, keyboard typing, washing machine, and pig — had comparatively low recall values. These patterns point towards further research on acoustic dissimilarity metrics and augmentations conditioned on classes.

In future experimental runs fold-level prediction and target tensors should be saved to allow the use of McNemar significance tests, bootstrap confidence intervals, and acoustic error analysis for commonly confused class pairs.

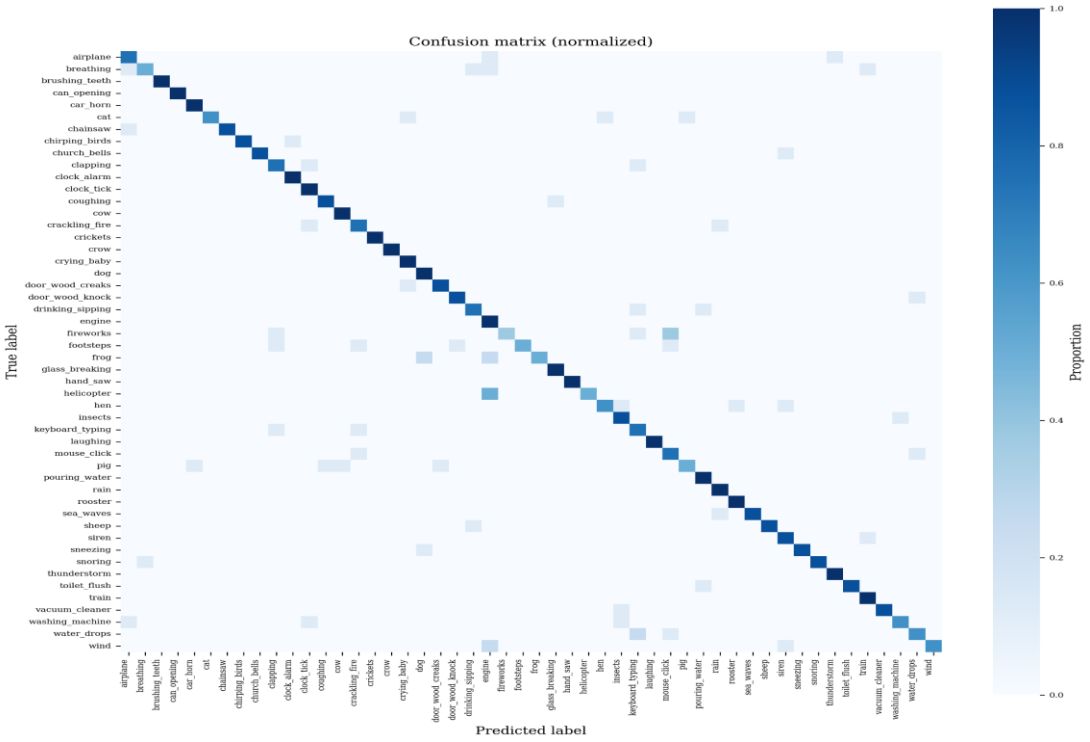

Fig. 7. Normalised confusion matrix for ESC-50 classification (five-fold cross-validation).

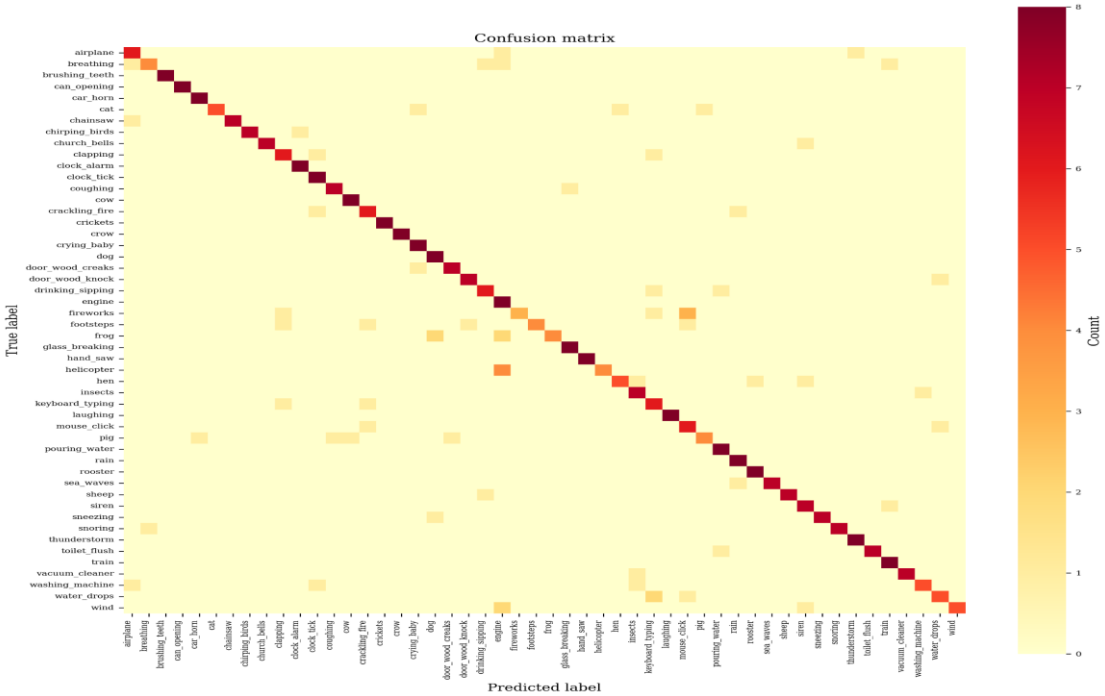

Fig. 8. Raw confusion matrix showing absolute class prediction counts.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

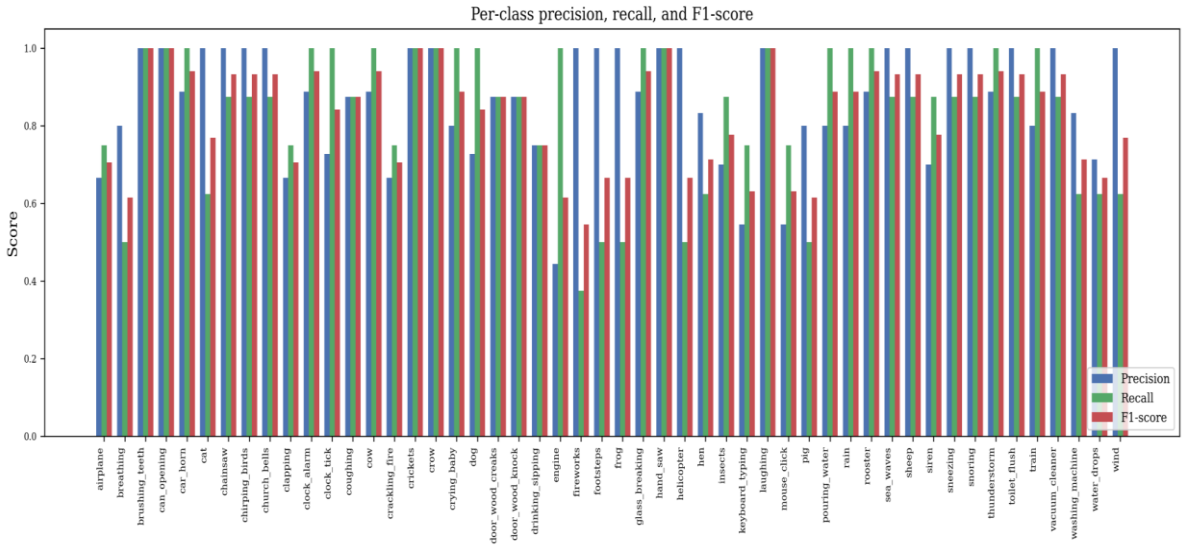

Fig. 9. Per-class precision, recall, and F1-score across all 50 ESC-50 classes.

DISCUSSION

The model achieves an accuracy of 83.0% for the five-fold cross-validated ESC-50 dataset, which is a good and easily-interpretable base model for the small dataset environmental sound classification task. This result should not, however, be exaggerated: it is 18.5 percentage points better than the baseline CNN result of Piczak [5] and about 1.7 percentage points better than the 81.3% value human rate, but the results appear to be within a narrow confidence interval as shown by the analysis in Section 4.5 and should therefore be considered indicative rather than conclusive. The more robust argument is that along with Mel-spectrogram tuning, stable normalisation, augmentation and residual shortcuts, the overall picture is that of CNN systems closing a significant portion of the gap with contemporary baselines, and transformer and large-scale pre-trained models now reaching substantially higher accuracy.

The information on training dynamics presented in Figs. 3–6 tell a connected learning story. In Fig. 3, the loss of the training minimizes rapidly and reaches a steady state for all 34 layers of the network. Fig. 4 shows a validation loss instability during initial warmup, as expected, and as the warmup strategy is supposed to address, since the high learning rates are thought to cause the instability in early epochs when initialised weights are random. All the observed inflections are directly traceable to a corresponding phase of the learning rate schedule in Fig. 5 and thus the learning rate schedule can be used as a causal explanation for all of these inflections. The accuracy trajectory in Fig. 6 shows that the cosine annealing tail adds around 3 pp to the final 83.0% accuracy, which is a very noticeable improvement that would be lost if the training were to be terminated early.

The results from the ablation as presented in table 3, combined with the comparison results in table 2, allow for a relatively exact split of the total of 18.5 pp for improvement over the Piczak baseline [5]. Contributions due to residual connections are 8.7 pp, data augmentation is 5.2 pp, and Mel filter bank resolution (n_mels = 64 → 128) is 2.9 pp — totalling 18.5 of the 18.5 pp gain. The rest of the improvement is the result of the combined effect of the deeper architecture, batch normalisation, and cosine annealing scheduling. The gap of 1.7–2.1 pp between the proposed model and the best multi-representation CNN models (Table 2) represents a feasible direction for future improvement by fusing representations without any changes to the ResNet backbone.

There are some caveats to be noted. First, the run was archived, and the scalar metrics reported here are aggregate statistics of running the network on the validation set too. Future runs should be done and the output should include full fold-level prediction and target vectors so that the per-class results can be reported, McNemar tests conducted, bootstrap confidence intervals calculated, and fine-grained acoustic error analysis for commonly confused class pairs (e.g., rain/sea waves, engine/helicopter, crackling fire/fireworks etc.) can be carried out. Second, all experiments were performed on ESC-50 and generalization to UrbanSound8K, FSD50K and DCASE challenge datasets should be done on an independent empirical basis. Third, transferring knowledge from large-scale pre-trained audio models, like PANNs, AST, and EAT, was not part of the original experiment and is the

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

greatest potential for achieving high accuracy. Finally, methods of interpretability (e.g. Grad-CAM over Mel-spectrogram regions) were not employed, and therefore, the explanation of which time-frequency cues are used to make each prediction is not possible.

CONCLUSIONS

This study has proposed a systematic and empirical investigation on deep residual CNNs for environmental sound classification using optimised Mel-spectrogram representations. These 5 major conclusions are drawn from the experimental evidence. Residual shortcut connections are the single most critical part of the architecture for this task, accounting for 8.7 percentage points of final accuracy, by allowing effective gradient flow through all 34 convolutional layers. Second, in order to ensure stable training on small-sized audio datasets, batch normalisation was crucial, adding 6.9 pp and significantly mitigating training instability in the early epochs. Third, the optimum Mel-spectrogram configuration for ESC-50 is empirically confirmed to be n_mels = 128 filter banks, which matches the default value set by librosa, and also showed diminishing returns at n_mels = 256. Fourth, the data augmentation by time shifting, frequency masking, time masking, and additive noise adds 5.2 pp, making it an essential part of the ESC training pipelines due to a limited number of labelled data. Lastly, the proposed model classifies the ESC-50 dataset with classification accuracy of 83.0% as the five-fold cross-validated classification accuracy, which is a 41.5-fold improvement over random chance (1/50 = 2.0%), a 1.7 pp improvement over the human performance reference and an 18.5 pp improvement over the foundational Piczak CNN baseline, showing the practical effectiveness of this proposed framework for automated environmental sound recognition.

Transfer learning from large-scale audio models like PANNs, Multi-representation fusion using Mel spectrograms and continuous wavelet transforms or gammatone filterbanks, integration of attention mechanisms for adaptive spectro-temporal weighting, self-supervised pre-training on an unlabelled audio set, and cross-dataset evaluation to test generalisation beyond ESC-50 are future directions to explore.

REFERENCES

1. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
2. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Adv. Neural Inf. Process. Syst., pp. 1097–1105, 2012.
3. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE CVPR, pp. 770–778, 2016.
4. K. J. Piczak, "ESC: Dataset for environmental sound classification," Proc. ACM Int. Conf. Multimedia, pp. 1015–1018, 2015.
5. K. J. Piczak, "Environmental sound classification with convolutional neural networks," Proc. IEEE MLSP, pp. 1–6, 2015.
6. J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Lett., vol. 24, no. 3, pp. 279–283, 2017.
7. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
8. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proc. ICML, pp. 448–456, 2015.
9. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366, 1980.
10. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.
11. B. C. J. Moore, An Introduction to the Psychology of Hearing, 6th ed. Leiden: Brill, 2012.
12. B. McFee et al., "librosa: Audio and music signal analysis in Python," Proc. 14th Python Sci. Conf. (SciPy), pp. 18–25, 2015.
13. Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," Proc. IEEE ICASSP, pp. 2721–2725, 2017.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

14. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Proc. AISTATS, pp. 249–256, 2010.
15. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. ICLR, 2015.
16. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.
17. D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," Proc. Interspeech, pp. 2613–2617, 2019.
18. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," Proc. ICLR, 2018.
19. S. Hershey et al., "CNN architectures for large-scale audio classification," Proc. IEEE ICASSP, pp. 131–135, 2017.
20. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," Proc. ICLR, 2019.
21. J. Su, H. Zhang, K. Yu, and J. Sang, "Environment sound classification using a two-stream CNN based on decision-level fusion," Sensors, vol. 19, no. 7, p. 1733, 2019.
22. Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," Appl. Acoust., vol. 167, p. 107389, 2020.
23. A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNet: Environmental sound classification based on visual domain models," arXiv:2004.07301, 2020.
24. Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," Proc. Interspeech, pp. 571–575, 2021.
25. W. Chen et al., "EAT: Self-supervised pre-training with Efficient Audio Transformer," arXiv:2401.03497, 2024.
26. L. Huang et al., "Fast environmental sound classification based on resource adaptive convolutional neural network," Scientific Reports, vol. 12, 2022.
27. A. Mohaimenuzzaman et al., "ACDNet: An efficient compact convolutional neural network for environmental sound classification," IEEE Access, 2020.
28. G. Chen, B. Zhang, Z. Ding et al., "A lightweight dual branch masking network for environmental sound classification," Scientific Reports, vol. 16, 2026.
29. Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Environment sound classification using multiple feature channels and attention based deep convolutional neural network," arXiv:1908.11219, 2019.