Gümüşhane University Journal of Science and Technology

GUFBD / *GUJS* (2026) 16(1): 142-159 doi: 10.17714/gumusfenbil.1797174 Research Article

ForestX-Net: a novel acoustics events recognition model for the forest environment using hybrid deep learning architecture

ForestX-Net: orman ortamı için hibrit derin öğrenme mimarisi kullanan yeni bir akustik olay tanıma modeli

Yunus KORKMAZ*

Dicle University, Faculty of Engineering, Department of Computer Engineering, 21200, Diyarbakir, Turkey

• Received: 04.10.2025 • Accepted: 05.02.2026

Abstract

Recognizing the acoustics events involves the identification and classification of auditory objects within various environments. For the forest ecosystems, it plays a vital role in domains such as monitoring biodiversity, detecting illegal activities, managing environmental threats and wildlife preservation. In this study, a novel framework which is named as ForestX-Net was proposed to classify forest acoustics events using a hybrid deep learning approach. The dataset comprises 10 distinct sound classes which are fire, rain, thunderstorm, helicopter, axe, chainsaw, gunshot, footstep, frog, and wolf howl, with 74 samples per class, recorded under authentic forest conditions. Spectrogram representations of these audio signals were extracted and employed as inputs to a pre-trained ResNet-18 model. Feature embeddings from ResNet-18 yielded a 740x512 feature matrix, which was subsequently utilized as input to a Multilayer Perceptron (MLP). The proposed architecture achieved an exact test accuracy of 92.57%, demonstrating its effectiveness in distinguishing acoustically diverse sound events.

Keywords: Deep learning, Forest acoustics, Machine learning, Sound classification, Transfer learning

Öz

Akustik olayların tanınması, farklı ortamlardaki işitsel nesnelerin belirlenmesini ve sınıflandırılmasını içermektedir. Orman ekosistemleri için bu durum; biyolojik çeşitliliğin izlenmesi, yasa dışı faaliyetlerin tespiti, çevresel tehditlerin yönetimi ve yaban hayatının korunması gibi alanlarda hayati bir rol oynamaktadır. Bu çalışmada, orman akustik olaylarını hibrit bir derin öğrenme yaklaşımı ile sınıflandırmak için ForestX-Net adlı yeni bir çerçeve önerilmiştir. Veri kümesi; yangın, yağmur, fırtına, helikopter, balta, motorlu testere, silah sesi, ayak sesi, kurbağa ve kurt uluması olmak üzere 10 farklı ses sınıfından oluşmaktadır ve her sınıfta gerçek orman koşullarında kaydedilmiş 74 örnek bulunmaktadır. Bu ses sinyallerinin spektrogram görüntüleri çıkarılmış ve önceden eğitilmiş ResNet-18 modeline girdi olarak verilmiştir. ResNet-18'den elde edilen öznitelik gömmeleri 740x512 boyutunda bir öznitelik matrisi üretmiş ve bu matris Çok Katmanlı Algılayıcı (MLP) modeline giriş olarak kullanılmıştır. Önerilen mimari %92,57 doğruluk oranı ile test başarısı göstermiş ve akustik açıdan farklı ses olaylarını ayırt etmedeki etkinliğini ortaya koymuştur.

Anahtar kelimeler: *Derin öğrenme, Orman akustik, Makine öğrenmesi, Ses sınıflandırma, Transfer öğrenme*

1. Introduction

Acoustics Event Recognition (AER) in natural environments has emerged as a critical research area with wide-ranging applications in ecological monitoring, environmental management, and conservation efforts (Simonović et al., 2021). Recognizing and classifying these sounds can provide valuable insights into monitoring biodiversity (Han & Peng, 2024), identifying illegal activities, and addressing critical environmental threats. As forests are home to diverse species and habitats, AER allows researchers to non-invasively study wildlife behaviors, detect endangered species, and monitor ecosystem health over time.

One of the most pressing challenges in forest management is the detection of illegal activities, including unauthorized logging, hunting, and human intrusion. Acoustic analysis can act as an early warning system by

*Yunus KORKMAZ; yunus.korkmaz@dicle.edu.tr

identifying sounds such as chainsaws, axes, and gunshots, which often accompany such activities (Sun et al., 2024). In addition, AER aids in managing environmental threats such as wildfires and extreme weather events. Sounds like fire crackling, thunderclaps, or helicopters during emergency responses can be classified to inform decision-makers and enable rapid intervention. Furthermore, wildlife preservation relies heavily on detecting and analyzing animal vocalizations, which help assess species diversity, population dynamics, and changes in habitat conditions. Given the complexity of forest soundscapes (characterized by overlapping signals, background noise, and diverse frequencies) traditional sound classification methods often fall short. Deep learning has emerged as a transformative solution for such challenges, offering state-of-the-art accuracy and scalability in sound recognition tasks (Tripathi & Mishra, 2021; Constantini et al., 2022; Tripathi & Paul, 2022; Mehrish et al., 2023; Fava et al., 2024; Wu et al., 2024; Aslam et al., 2024; Wang et al., 2024; Chang et al., 2025; Arafath & Routray, 2025).

In this study, ForestX-Net, a novel cross-deep learning framework was proposed for recognizing and classifying acoustics events in forest environments. The term "X" in ForestX-Net signifies the cross-network architecture that integrates two deep learning techniques: feature extraction using a pre-trained ResNet-18 network and classification using a Multilayer Perceptron (MLP). ResNet-18, a well-established deep residual neural network, was employed to extract high-level feature embeddings from spectrogram images of forest sounds. Spectrogram images have been used in many scientific problems by the researchers (Saravanan et al., 2018; Mushtaq et al., 2021; Dissanayake et al., 2023; Abraham et al., 2023; Yi et al., 2024; Tang & Hu, 2024; Fan et al., 2024; Zhang et al., 2025). These embeddings are then fed into an MLP network, which effectively classifies the features into predefined categories. This cross-network approach leverages the strengths of both ResNet-18 and MLP, addressing the challenges of sound classification in forest environments.

The ResNet-18 architecture has gained prominence in deep learning due to its ability to handle deep feature extraction through residual learning (Zhu et al., 2021; Xiao et al., 2022; Manivannan, 2022; Xiang et al., 2023; Yeh et al., 2023). Its use as a feature extractor in this study allows for the generation of robust, high-dimensional feature embeddings from spectrogram representations of acoustics events. On the other hand, the Multilayer Perceptron (MLP) serves as a powerful classifier for the extracted features (Khishe et al., 2018; Sharma et al., 2022; Krishina & Kokil, 2023; Panimalar et al., 2025; Gao et al., 2025). By utilizing the embeddings generated by ResNet-18 as input, the MLP network in ForestX-Net effectively distinguishes between 10 acoustics event classes: fire, rain, thunderstorm, helicopter, axe, chainsaw, gunshot, footstep, frog, and wolf howl. The proposed ForestX-Net framework achieves a remarkable test accuracy of 92.57% on a dataset containing 740 spectrogram images, with 74 samples per class. This performance highlights the effectiveness of combining ResNet-18's feature extraction capabilities with MLP's classification power in addressing the unique challenges of forest sound classification.

1.1. Literature works

The ability to automatically identify and classify sounds within forests facilitates the monitoring of biodiversity, detection of illegal activities, management of environmental threats, and preservation of wildlife. Recent studies have emphasized the significance of acoustic monitoring in assessing biodiversity. For instance, the development of the FSC22 dataset, comprising 2,025 sound clips across 27 acoustic classes, has provided a benchmark for forest environmental sound classification, aiding researchers in developing more accurate models (Bandara et al., 2023). Additionally, the application of deep learning methods for forest acoustics classification have been investigated to balance accuracy and model complexity, facilitating implementation on resource-limited edge devices (Paranayapa et al., 2024). In the realm of wildlife preservation, passive acoustic monitoring has proven invaluable for studying animal behavior and habitat use. For example, the classification of animal sounds in hyperdiverse rainforests using Convolutional Neural Networks (CNNs) has facilitated the detection of species presence and activity patterns, contributing to more effective conservation strategies (Sun et al., 2021). Furthermore, the integration of Internet of Things (IoT) devices for forest sound classification has been explored, with each node's status transmitted to a gateway at the forest base, enhancing real-time monitoring capabilities (Peng et al., 2023; Ayankoso et al., 2024). The development of physics-based models to predict the acoustic detection distance of terrestrial autonomous recording units over the diel cycle and across seasons has provided insights into the effective deployment of acoustic sensors in diverse forest environments (Haupert et al., 2022). Recent studies have demonstrated the effectiveness of ResNet-18 in various sound classification domains. For instance, fine-tuning ResNet-18 for audio classification tasks has yielded high accuracy on datasets like ESC-50, showcasing its adaptability to different audio datasets (Zhiqing

et al., 2024). In the context of environmental sound recognition, ResNet-18 has been employed to classify urban sounds, contributing to the development of smart city applications (Nogueira et al., 2022). The versatility of ResNet-18 extends beyond audio classification, with applications in spectral classification tasks such as multispectral remote sensing images, indicating its robustness in handling various types of spectral data (Zhao et al., 2022). Moreover, the integration of attention mechanisms with ResNet-18 has been explored to enhance its performance in acoustic scene classification. For example, an attention-based ResNet-18 model has been proposed for the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, demonstrating improved accuracy in distinguishing between different acoustic scenes (Aryal & Lee, 2020).

Multilayer Perceptrons (MLPs) have been effectively utilized in sound classification tasks, often serving as classifiers that process features extracted from audio signals. In AER tasks, MLPs have been used to process features derived from spectrograms, achieving competitive accuracy levels (Shanthakumar et al., 2020). The Paired Inverse Pyramid Structure MLP Network (PIPMN) is one such architecture that has demonstrated high accuracy in environmental sound classification tasks, achieving 96% accuracy on the UrbanSound8K dataset (Chen et al., 2022). Also, deep MLPs have been applied to dimensional speech emotion recognition, outperforming modern deep learning architectures like Long Short-Term Memory (LSTM) networks and CNNs in certain scenarios (Atmaja & Akagi, 2020). Additionally, MLPs have been compared with Support Vector Machines (SVMs) in speech and song emotion recognition tasks, demonstrating comparable performance and highlighting the versatility of MLPs in handling various audio classification challenges (Javaheri, 2021). Furthermore, MLP-ASR models introduce sequence-length agnostic all-MLP architectures, showcasing the potential of MLPs in processing sequential audio data (Orosoo et al., 2025). These advancements underscore the significance of MLPs in the development of robust sound classification systems.

In environmental sound classification, machine learning models have been trained to distinguish between different types of sounds, such as urban noises, animal calls, and natural phenomena. A study introduced a deep learning-based approach for classifying poultry audio signals, incorporating a custom Burn Layer to enhance model robustness (Hassan et al., 2024). Another research presented a deep learning model for multi-label sound classification that can be deployed in real-world scenarios on edge devices (Goulão et al., 2024). A comprehensive survey of audio classification using deep learning provides insights into various models applied for such tasks (Zaman et al., 2023). Additionally, a study on deep learning-based lung sound analysis highlights the significance of artificial intelligence in medical sound classification (Huang et al., 2023). Researchers have also developed frameworks for animal sound classification with feature optimization, enhancing the accuracy of species identification (Akbal et al., 2022). Moreover, studies on decoding birdsong have utilized machine learning to gain new insights into the complexity of bird communication (Nanni et al., 2020). Another study introduced SoundCLR, a supervised contrastive learning method for environmental sound classification, achieving state-of-the-art performance (Nasiri & Hu, 2021). A research proposed an end-to-end approach for environmental sound classification based on a 1D CNN (Abdoli, 2019). In another study, authors designed a deep learning-driven system to analyze acoustic signals generated by industrial machinery (Yurdakul & Tasdemir, 2023). Another research focused on abnormal acoustics detection utilizing audio representations pre-trained with machine ID-based contrastive learning (Guan et al., 2023).

2. Theoretical background

The audio signals are mostly not periodic signals. They can be found as non-stationary signal in natural acoustic environments. They require to be modelled as sine waves, cosine waves or combination of both. To achieve such a modelling, well-known Fast Fourier Transform (FFT) can be applied to raw audio signal (Cooley & Tukey, 1965).

The log-mel spectrogram is a kind of image-based representation for sound signals. Three-dimensional (3D) spectrograms were then used as input for pre-defined ResNet-18 deep neural network. Figure 1 shows sample spectrograms extracted from each class in used dataset. ResNet-18 was formed using Convolutional Neural Networks (CNNs) (He et al., 2016). The ResNet-18's ability to extract high-level features from Log-Mel spectrograms makes it suitable for this study. The ResNet-18 network architecture was visualized by Figure 2. As in Figure 2, ResNet-18 employs two distinct filter sizes, 7×7 and 3×3, which are strategically integrated to optimize the model's training and feature representation capabilities. Once convolution steps were done, a pooling operation is used with the size of 1x1x512. There is also fully connected layer having size of 1000 at the end of the model. Softmax function is used at the end of ResNet-18 to convert the network's final outputs

into probabilities. MLPs are particularly effective in classification tasks because of their fully connected architecture, which allows them to learn intricate patterns in the input data as it was given by Figure 3. In Figure 3, X1-X4, H1-H8 and "out" indicate the neurons and W1-W3 represents the network weights. The MLP serves as the classifier following the feature extraction step (via ResNet-18) in this work. The MLP processes the feature embeddings generated by ResNet-18 and assigns them to specific sound classes.

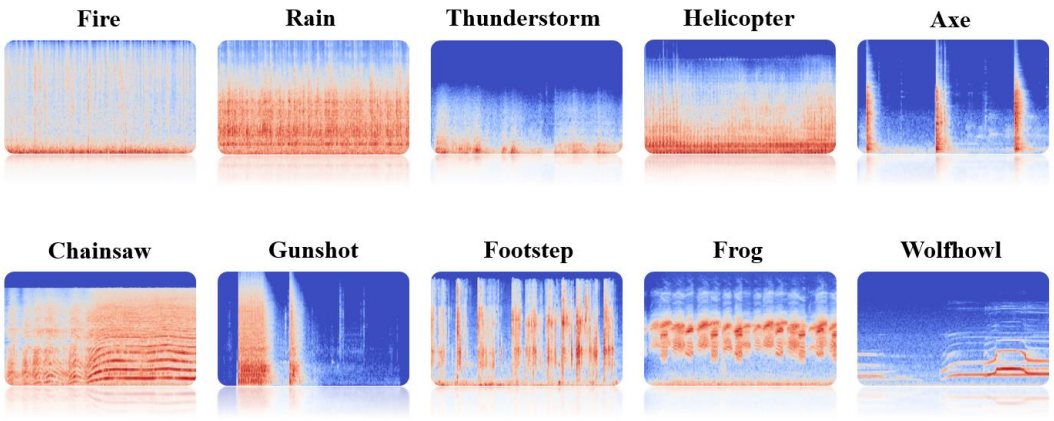


Figure 1. Spectrogram examples in dataset.

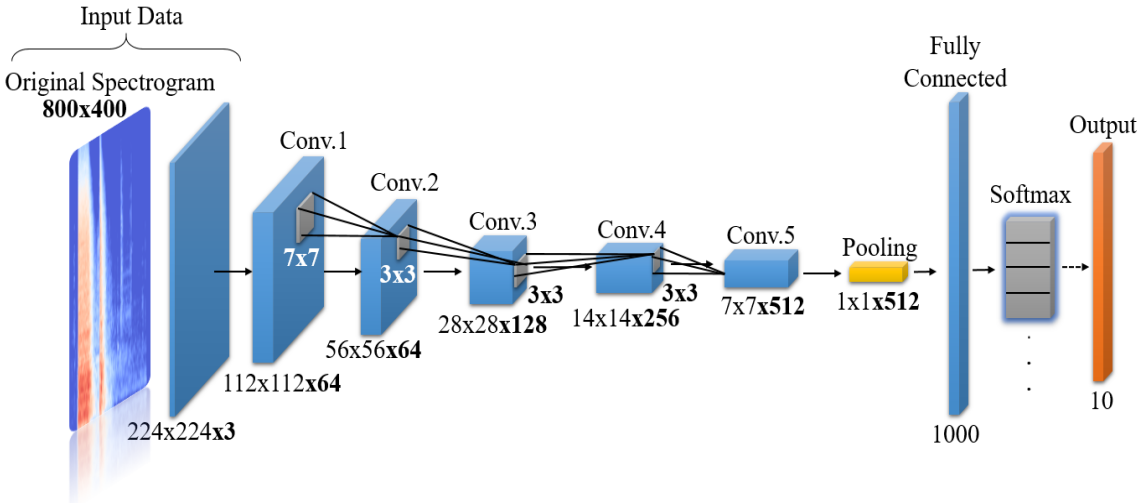


Figure 2. The ResNet-18 deep neural network architecture.

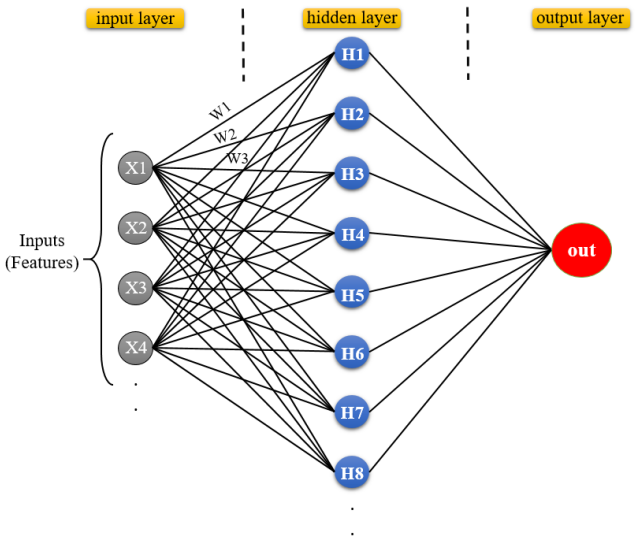


Figure 3. The structure of Multilayer Perceptron neural network.

3. Proposed model: ForestX-Net

The proposed sound recognition system for forest environments can roughly divided into three steps which are log-mel spectrogram extraction, ResNet-18 feature extraction and MLP-based multi-class classification. The final fully connected classification layer of ResNet-18 was removed so that the network outputs a 512-dimensional embedding. This embedding is then fed into an external MLP classifier. The proposed model was named as ForestX-Net ("X" means cross deep networks). General outline of the ForestX-Net was provided by Figure 4. In step 1, log-mel spectrogram extraction was directly applied to audio recording samples to obtain totally 740 spectrogram images (74 for each of 10 classes).

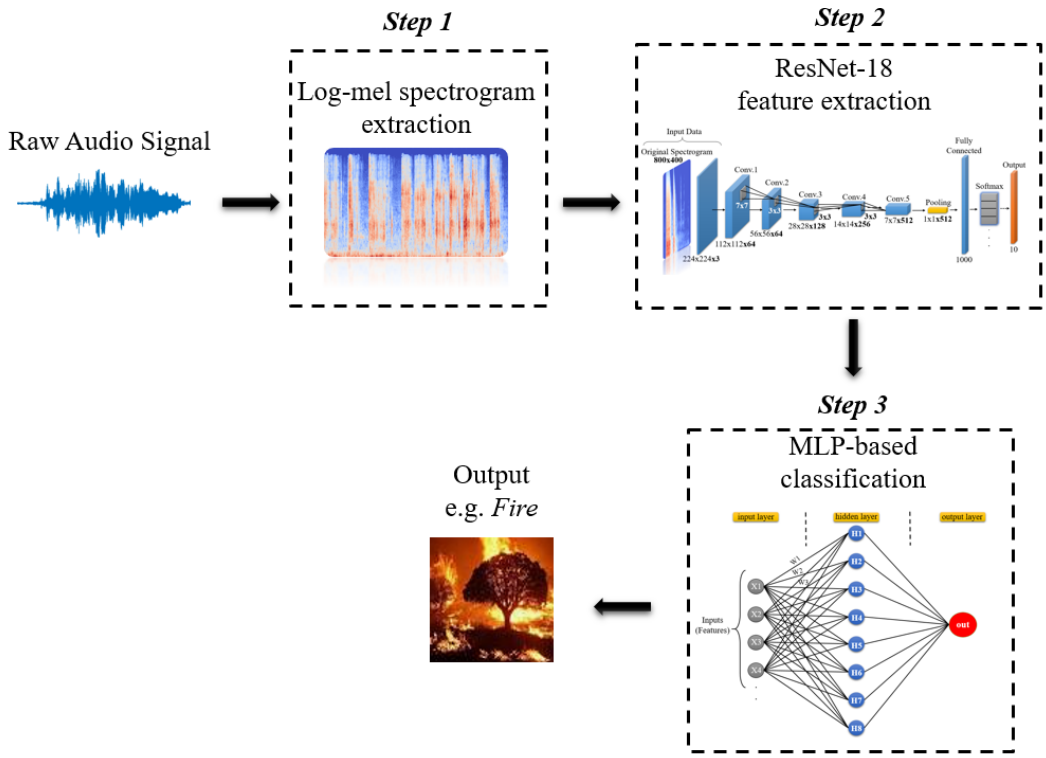


Figure 4. General outline of the proposed ForestX-Net.

Table 1. Hyper parameters for generation of log-mel spectrogram images.

Parameter	Value
Sampling frequency	44100 Hertz
FFT size	2048
Number of mel bands	128
Hop length (samples)	1024

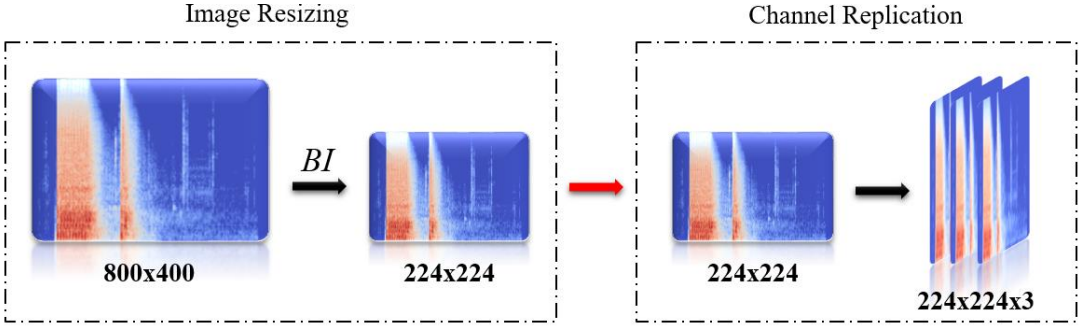


Figure 5. Image input preparation for the ResNet-18 feature extractor.

Spectrogram extraction process requires some hyper-paramaters for mathematical calculations. Table 1 shows all the input parameters used for spectrogram generation procedure. After spectrograms were obtained with the size of 800x400, all images were re-sized to 224x224x3 in order to be compatible with input shape of ResNet-18 network. This process was done using two individual steps which are image resizing and channel replication as it was represented by Figure 5. For second step in Figure 4, ResNet-18 was fine-tuned for the forest sound recognition problem. ResNet-18 was originally trained on ImageNet dataset (Deng et al., 2009). It can be fine-tuned for specialized datasets, significantly reducing the computational cost compared to training a model from scratch (Weiss et al., 2016). Fine-tuning parameters for ResNet-18 was given by Table 2.

Table 2. Parameters before and after fine-tunings (FT) of pre-trained ResNet-18.

Layer	Value Before FT	Value After FT	Layers, Con. & Params
Fully Connected	*WLRFactor*:1	*WLRFactor*:10	71, 78 & 11 M
	BLRFactor:1	*BLRFactor*:10	
Output	*OutputSize*: 1000	*OutputSize*: 10	

For the fine-tuning process of ResNet-18, SGDM (stochastic gradient descent with momentum) optimizer using a momentum coefficient of 0.9 was employed. The initial learning rate was set to $1x10^{-4}$. The network was trained for 30 epochs with a mini-batch size of 32. These settings follow common practices in transfer-learning based feature extraction studies and were selected to balance stability and overfitting control.

In addition to the ResNet-18 backbone used in ForestX-Net, two widely adopted pretrained audio models which are YAMNet and VGGish (Hershey et al., 2017), were incorporated into the experimental design as comparative baselines. YAMNet is based on MobileNet-V1 and trained on AudioSet, producing a 1024-dimensional embedding for each audio clip. VGGish, on the other hand, outputs a 128-dimensional embedding extracted from log-mel spectrograms. In both cases, our pipeline replaced their original classification layers and extracted only the feature embeddings. These embeddings were subsequently passed to the same external MLP classifier used in ForestX-Net, ensuring a controlled comparison where only the feature extractor differs. This evaluation aimed to understand whether audio-specialized models pretrained on large-scale sound corpora outperform a fine-tuned ResNet-18 under the constraints of the FSC22 dataset.

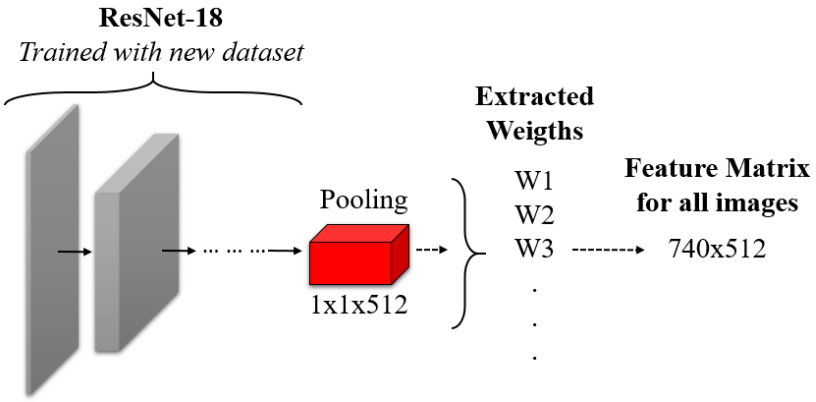


Figure 6. Feature embedding extraction scheme.

Table 3. Hyper-parameters for the MLP classifier.

Parameter	Value
Batch size	256
Hidden layer size	300
Solver	Adam
Activation function	tanh
Iteration	230

The ADAM (Adaptive Moment Estimation) optimizer (Kingma & Ba, 2014) was used as solver function. The optimizer updates model parameters based on estimates of the first moment (mean) and the second moment (uncentered variance) of the gradients. The parameter updates were shown by Equatin 3, Equation 4, Equation 5 and Equation 6.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{3}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{4}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{5}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \tag{6}$$

where m_t and v_t are the biased first and second moment estimates, $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates, θ_t represents the parameters being updated, g_t is the gradient at time t, η is the learn rate, ϵ is a small value introduced to ensure computational precision. Hyperparameters β_1 and β_2 control the rates at which the moving averages decay and are set to 0.9 and 0.999, respectively. Also, the "tanh" activation function which was employed within the Multilayer Perceptron (MLP) is defined by Equation 7. Output interval of the "tanh" function is -1 to 1.

$$f(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{7}$$

4. Experimental results

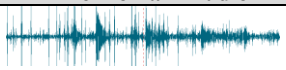
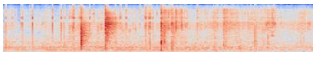
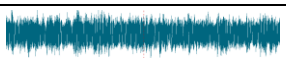
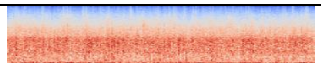
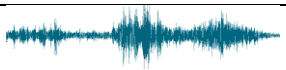
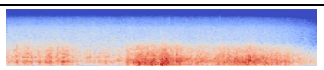
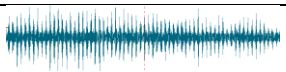
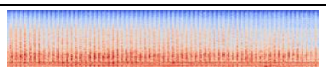
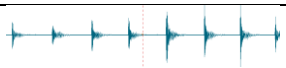
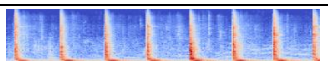
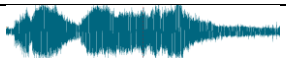
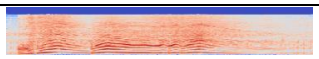
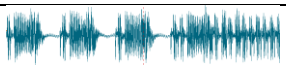
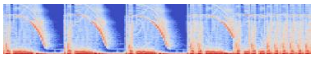
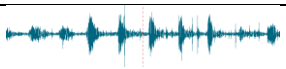
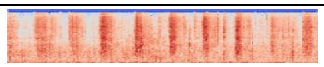
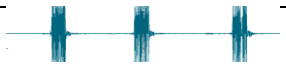
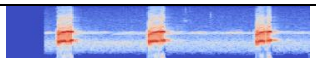
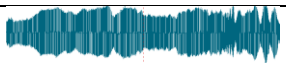
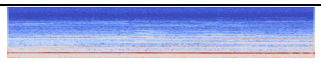
In order to see the success of the proposed model, dataset has been tested under different scenarios. Five types of tests were used to evaluate the proposed model. All test scenarios were shown by Table 4. For the T1, T2 and ForestX-Net, ResNet-18 deep neural network was used as feature extractor. For the T3 and T4, spectrogram images were directly used as features for the pre-trained ResNet-18 and AlexNet. Both of T3 and T4 can be seen as applications of transfer learning.

Table 4. Five test scenarios for a comperative analysis.

Scenario	Feature	Classifier
T1	740x512 from ResNet-18	The ML without PCA (KNN classifier)
T2	740x512 from ResNet-18	The ML with PCA (SVM classifier)
T3	Log-mel spectrogram images	ResNet-18 (a pre-trained deep network as classifier)
T4	Log-mel spectrogram images	AlexNet (a pre-trained deep network as classifier)
ForestX-Net	740x512 from ResNet-18	Multilayer Perceptron (MLP)

As mentioned before, a total of 740 spectrogram images were used for all training phase and only 20% of 740 was used for tests. Dataset in this work is a 10-class subset of FSC22 dataset which was recently published as open dataset (Bandara et al., 2023). The reason of selecting 10-class subset of FSC22 is that these classes represent key categories relevant to forest monitoring, including natural events (fire, rain, thunderstorm), biological activity (frog, wolf howl), and potential human-induced threats (axe, chainsaw, gunshot, footstep). This selection ensures that the model focuses on acoustic events that are directly associated with biodiversity tracking, illegal activity detection, and environmental event monitoring. Samples from log-mel spectrogram of each classes were provided in Table 5. Time-Domain Audio shows the raw form of audio signals in dataset while spectrogram images of corresponding audio were represented in column of Frequency-Domain Audio. # of Sam. indicates the number of observations for each class in dataset.

Table 5. Structure of the dataset used for all test scenarios.

Class	Time-Domain Audio	Frequency-Domain Audio	Spec. size	# of Sam.
Fire			800x400	74
Rain			800x400	74
Thunderstorm			800x400	74
Helicopter			800x400	74
Axe			800x400	74
Chainsaw			800x400	74
Gunshot			800x400	74
Footstep			800x400	74
Frog			800x400	74
Wolfhowl			800x400	74

Sampling frequency of all forest audio samples are 44.1 KHz. “Spectrogram size” of all samples was reshaped as 224x224 and 227x227 for T3 and T4, respectively. Fine-tunings declared before by Table 2 was also applied to T1, T2, T3 and T4. The classification performances of T1, T2, T3, T4 and ForestX-Net were given in Table 6. All rates are out of hundred and they are all test accuracies.

Table 6. Classification results of experiments over forest sound dataset.

Experiments	Test Accuracy (%)
T1	82.4
T2	90.5
T3	90
T4	85.33
ForestX-Net	**92.57**

The proposed ForestX-Net with 92.57% test accuracy outperformed all other techiniques for the forest sound recognition task. This clearly shows the contribution of usage of ResNet-18 as feature extractor together with MLP classifier.

Table 7. Class-level precision, recall and F1 score of ForestX-Net.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Precision	0.94	0.93	0.88	1	0.92	0.94	0.93	0.875	0.93	0.93
Recall	1	0.93	1	0.79	0.8	1	0.93	0.93	0.93	0.93
F1 Score	0.97	0.93	0.94	0.88	0.86	0.97	0.93	0.9	0.93	0.93

In Table 7, sound-wise precision, recall and F1 score were provided to see the system’s success over each class. S1, S2, S3, S4, S5, S6, S7, S8, S9 and S10 refer to sounds of fire, rain, thunderstorm, helicopter, axe, chainsaw, gunshot, footstep, frog, and wolf howl, respectively. Average values for precision, recall and F1

score are the same, which is 0.93. It should not be forgotten that these values are for the best result which is 92.57% test accuracy obtained by ForestX-Net. Experiments of T1 and T2 has been carried with 10k cross validation folds to protect ML algorithm against overfitting. Receiver Operating Characteristic (ROC) curves of both T1 and T2 tests were shown by Figure 7 and Figure 8, respectively. Aslo, the ROC curve for the proposed ForestX-Net was shown by Figure 9. The Receiver Operating Characteristic curve is created by calculating the True Positive Rate (TPR) and False Positive Rate (FPR) over a range of threshold values, typically chosen at specific intervals. The curve is generated by plotting the TPR against the FPR for each threshold. TPR and FPR were calculated using Equation 8 and Equation 9, respectively, on the basis of confusion matrix given by Figure 10.

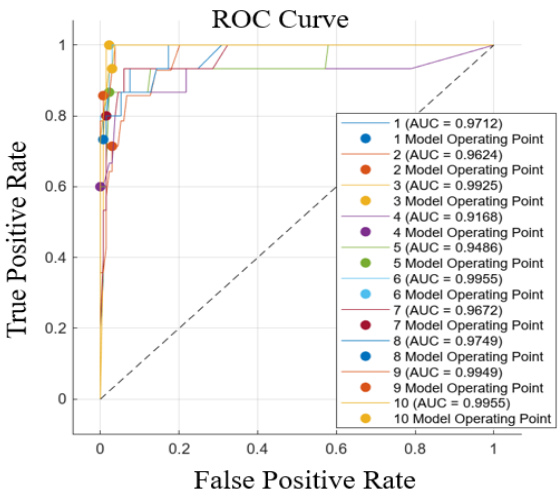


Figure 7. ROC curve for the experiment of T1.

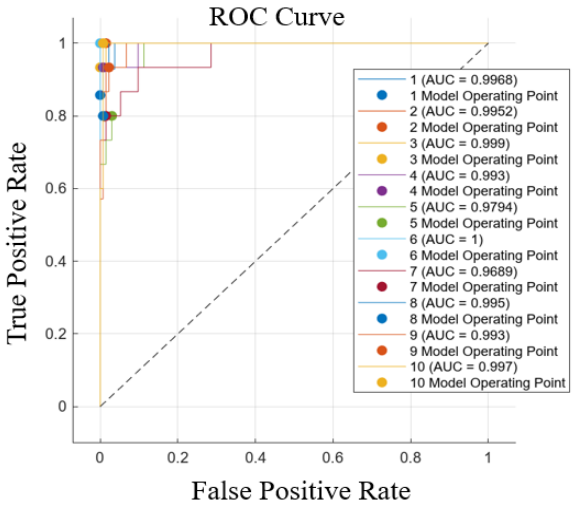


Figure 8. ROC curve for the experiment of T2.

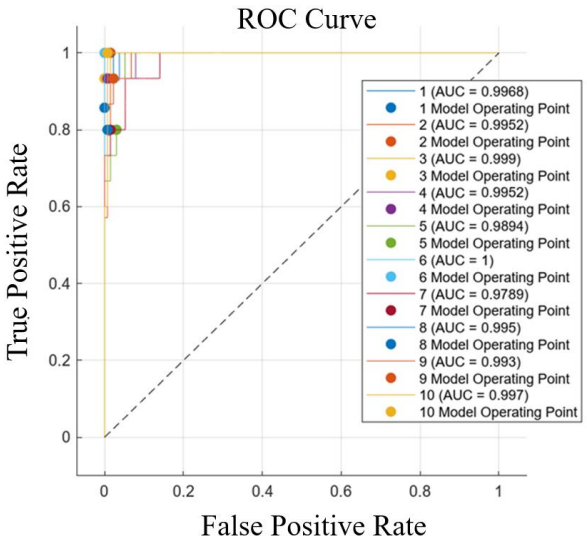


Figure 9. ROC curve for the proposed model ForestX-Net.

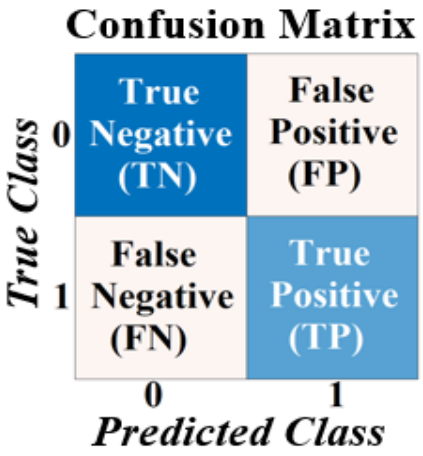


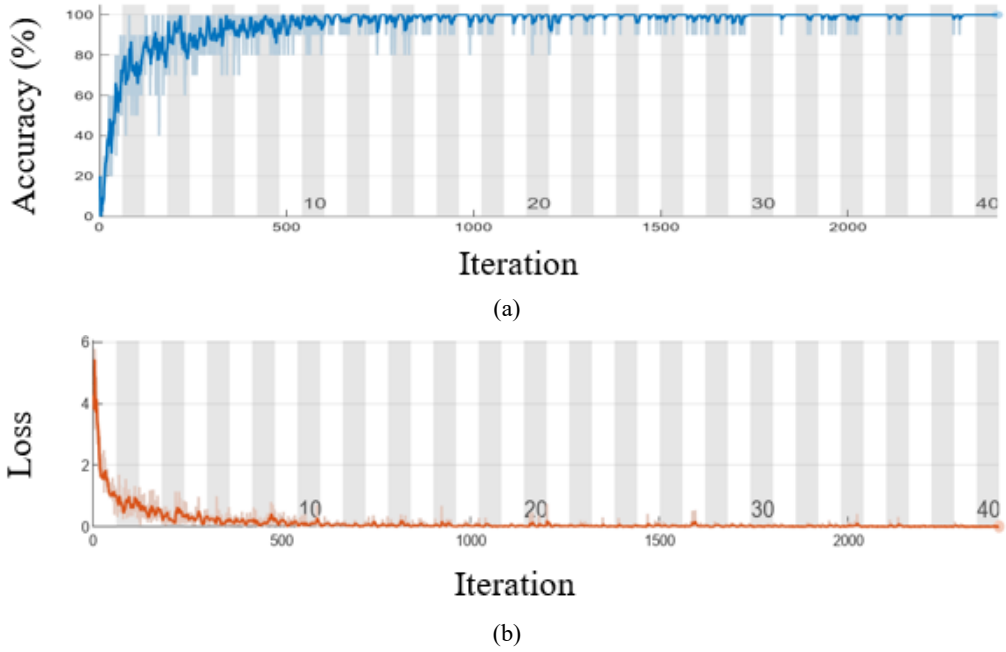
Figure 10. Confusion matrix indicating TN, FP, FN and TP.

$$TPR = \frac{TP}{TP+FN} \tag{8}$$

$$FPR = \frac{FP}{FP+TN} \tag{9}$$

Training and loss plots for both T3 (ResNet-18) and T4 (AlexNet) tests were represented by Figure 11 and Figure 12, respectively. It can clearly be inferred that ResNet-18 tracks better learning and loss curves than AlexNet. This is the main reason of chosing ResNet-18 for the proposed ForestX-Net. To ensure a robust

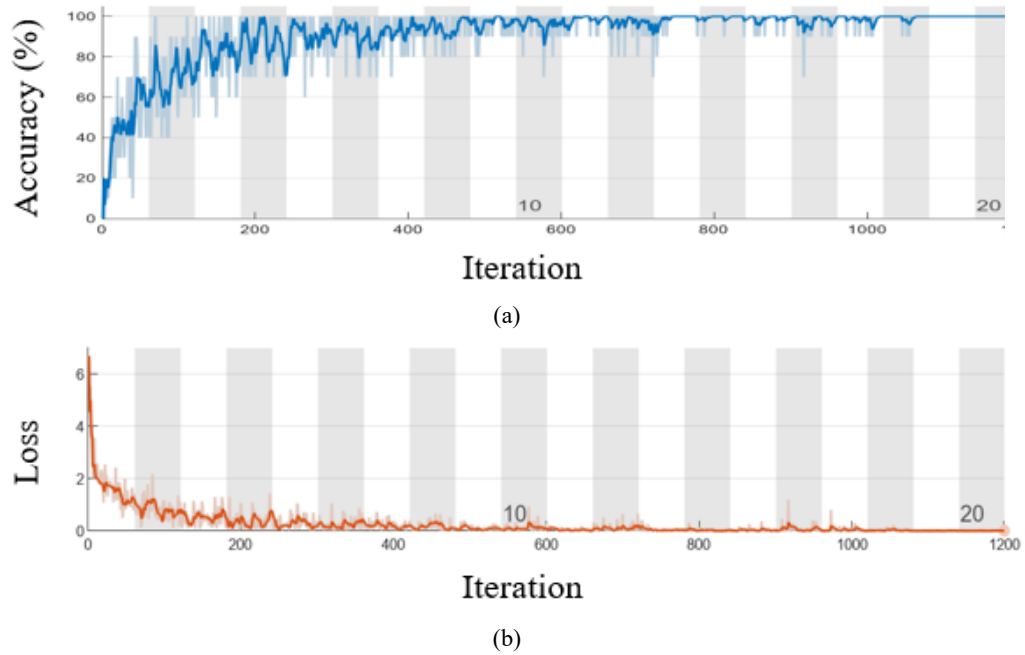
evaluation of the model's generalization ability, both k-fold cross-validation (with 5-fold and 10-fold) and a conventional train-test split approach were applied. The results indicate that the model achieved its highest accuracy (92.57%) when evaluated using a fixed 80-20 train-test split. Accuracies in the cross-validation classification results are 89.19% for 5-fold and 90.54% for 10-fold cross-validation, respectively. Comparison of the results was presented in Table 8.



(a)

(b)

Figure 11. Training (a) and loss (b) plots for the T3 test.



(a)

(b)

Figure 12. Training (a) and loss (b) plots for the T4 test.

Table 8. Classification performance of cross-validation and train-test split approaches over the ForestX-Net.

Approach	Accuracy (%)
Train-Test Split (80-20)	92.57
5-Fold Cross-Validation	89.19 ± 1.62
10-Fold Cross-Validation	90.54 ± 0.97

Instead of ResNet-18 block in pipeline of the proposed ForestX-Net, well-known YAMNet and VGGish (Hershey et al., 2017) pre-defined audio networks were used as feature extractors in order to compare the ForestX-Net to the mentioned networks in terms of network size and classification accuray. The results indicating number of parameters in million (M), network size in MB and test accuracy as percentage were provided by Table 9.

Table 9. Comparison of the ForestX-Net (ResNet-18 + MLP), YAMNet, and VGGish models.

Model	Number of Params (M)	Model Size (MB)	Test Accuracy (%)
ForestX-Net	~ 11.2	~ 42.7	92.57
YAMNet	~ 3.8	~ 14.8	91.21
VGGish	~ 6.1	~ 24.4	88.51

For YAMNet and VGGish, the tests were done under the same condition (same hyper parameter settings) with the ForestX-Net. The results demonstrated that ForestX-Net strikes an effective balance between model capacity and performance by using ResNet-18 block inside alternatively to other two pretrained networks. Confusion matrices of YAMNet and VGGish were shown by Table 10. Also ROC curves and class-wise AUC scores of YAMNet and VGGish related experimental settings were provided by Figure 13.

Table 10. Confusion matrix for the (a) ForestX-Net with YAMNet, (b) ForestX-Net with VGGish.

		Predicted Class									
True Class		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
	1	12			1		2				
	2		11				1	2			
	3			15							
	4				15						
	5					14		1			
	6	2				1	11	1			
	7							14		1	
	8		1						14		
	9									14	
	10										15

(a)

		Predicted Class									
True Class		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
	1	12			2		1				
	2		12				1	1			
	3			13	1		1				
	4	1			14						
	5					14		1			
	6	1					12		1		1
	7							14		1	
	8			1					13	1	
	9									14	
	10					2					13

(b)

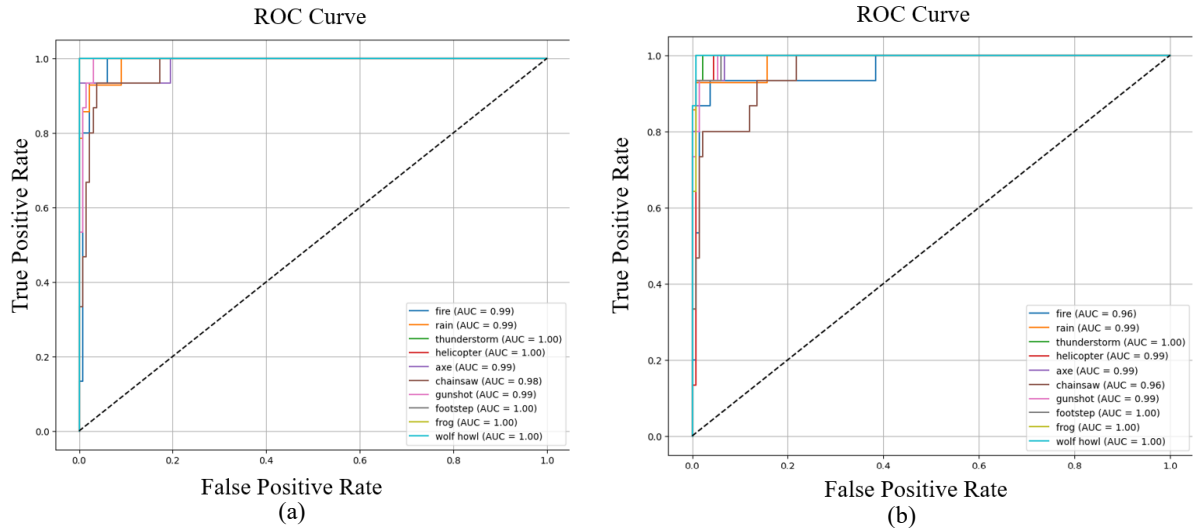


Figure 13. ROC curves & AUC scores of (a)YAMNet block (b)VGGish block in ForestX-Net.

Table 11. Confusion matrix for the (a) T1, (b) T2, (c) T3, (d) T4 and (e) ForestX-Net.

True Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10
1	13	1				1				
2	1	9	2	1		1				
3			15							
4		2	1	9	1					2
5					13		2			
6						14			1	
7		1			1		12			1
8	2				1	1		11		
9								1	12	1
10						1				14

(a)

True Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10
1	12				2					
2		14								
3			14						1	
4				14						1
5					12		2	1		
6						15				
7		1			1		12		1	
8		1			1			12	1	
9				1					14	
10										15

(b)

True Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10
1	15									
2		15								
3	2		13							
4				13				1		1
5			2		13					
6						12		2	1	
7							15			
8						1		13		1
9						1			14	
10					1	1		1		12

(c)

True Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10
1	14				1					
2		15								
3	1		12	2						
4			2	11	1					1
5			2		13					
6						11		1	3	
7		1					13	1		
8					2	2		10		1
9									15	
10					1					14

(d)

True Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10
1	15									
2	1	14								
3			15							
4		1	1	11		1				
5					12		1	2		
6						15				
7			1				13			
8					1			14		
9									14	1
10									1	14

(e)

5. Discussion

In this paper, a novel approach for the forest sound recognition was proposed to solve recognition problem in forest environments. The findings of the proposed model, ForestX-Net, were compared to different experiment scenarios which were mentioned as T1, T2, T3 and T4 aliases in Section 4, i.e. Results. Considering the Table 6, ForestX-Net reached to 92.57% test accuracy which is the best among rest of the experiments. Additionally, among all other classes, the ForestX-Net classified 15 fire sounds with 100% test accuracy. This also proves that the proposed model can be used to detect forest fires especially when the fire just started. By examining the Table 11 (e), it can be seen that the ForestX-Net reached 93.3% test accuracy by misclassfying only one images in each of classes of Rain, Footstep, Frog and Wolfhowl. The ForestX-Net's ability to make a well-discrimination between animal sounds and others showed that the proposed model is succesfull in monitoring biodiversity, wildlife preservation and determination of animal migration routes.

As observed in Table 7 and the confusion matrices in Table 11, certain classes such as helicopter (S4) and axe (S5) exhibited relatively lower recall (0.79 and 0.80, respectively). A detailed examination of the confusion matrices revealed that S4 (helicopter) was occasionally misclassified as S2 and S3, which is likely due to overlapping frequency patterns and similar temporal dynamics in their spectrograms. Similarly, S5 (axe) showed confusion with S6, possibly because both sounds shared sharp, percussive characteristics, leading to

spectrogram similarities. These misclassifications indicated that, while ForestX-Net effectively captured class-specific acoustic features overall, classes with acoustically similar profiles were more prone to confusion.

As shown by Table 9, both YAMNet and VGGish underperformed compared to the fine-tuned ResNet-18. This may be attributed to domain mismatch, as AudioSet recordings differ substantially from forest acoustics in terms of background noise profile, spectral density, and sound event duration. ResNet-18, when fine-tuned directly on FSC22 spectrograms, appears to adapt more effectively to the narrower frequency bands and temporal structures found in forest conditions. This justifies its selection as the core feature extractor in ForestX-Net.

Recognition performance of the proposed ForestX-Net has been compared to results of currently available studies (state-of-the-arts-*SOAs*) in literature. Table 12 shows all comparisions in terms of four cirteria which are feature, model, dataset and accuracy. As it can be seen from Table 12, limited studies using FSC22 are available in the literature. Because this study focused on only sounds in forest environment, other datasets like ESC-10, ESC-50 and UrbanSound8K were not prefered in this work as they include various sounds irrelevant to forest environment.

Table 12. Performance comparision of the ForestX-Net with state-of-the-arts (SOAs).

Author(s)	Feature	Model	Dataset	Accuracy (%)
(Xu & Chen, 2024)	Combination of 7 acoustic features & MFCC	SVC, KNN, DT, RF, ERF	Forest (FSC22)	~90 (with 10 classes)
(Tsalera et al., 2021)	Spectrograms	GoogleNet VGGish	Environmental Sounds	86.25 91.25
(Segarceanu et al., 2020)	MFCC, GMM	FFN-DFT	4 classes of Forest sounds	72.77 (chainsaw)
(Aytar et al., 2016)	Waveform	SoundNet	Environmental Sounds (10 cl.)	92.2
(Piczak, 2015)	Segmented Spectrograms	Simple CNN	UrbanSound8K	73.1 (LP) 73.7 (US)
The ForestX-Net (**proposed model**)	Spectrograms	ResNet-18 + MLP	Forest (FSC22)	**92.57**

One may clearly identify that the ForestX-Net outperformed many state-of-the-arts (SOAs) considering classification accuracy which is 92.57%. It is also important to note that while classes of “fire”, “chainsaw” and “thunderstorm” show perfect recall which is 1, certain classes like "helicopter" and “axe” exhibited a lower recall value compared to average as it can be derived from Table 7. Recall score of 1 means that the model has identified every instance of these sounds correctly. On the other hand, lower recall values for "helicopter" and “axe” classes indicate that, while the model correctly identified a high number of instances from this class, there were some "false negatives" (i.e., the model failed to identify some "helicopter" or “axe” sounds as such). A precision of 1 for "helicopter" class suggests that the model consistently predicts helicopter sounds correctly without incorrectly classifying other sounds as helicopter. The high F1 score values (such as 0.97 for "fire" and "chainsaw") indicate that ForestX-Net performs exceptionally well in these classes, achieving a strong balance between precision and recall. The ForestX-Net’s performance was also evaluated using cross-validation approach. There are differences in classification accuracies when using 5-fold-cross-validation, 10-fold-cross-validation and fixed train-test split size. One key reason for this difference is the limited dataset size.

Since FSC22 is a domain-specific dataset targeting forest acoustic monitoring, the number of existing studies using this dataset is currently very limited. For this reason, broader environmental sound benchmarks such as ESC-50 and UrbanSound8K were included to contextualize the performance of the proposed method with respect to widely used datasets in the sound classification literature.

The relatively small number of samples in FSC22 (74 per class) poses an intrinsic limitation for deep learning models, particularly those involving convolutional front-ends. The discrepancy observed between the fixed split accuracy (92.57%) and the cross-validation averages (89.19% and 90.54%) suggests sensitivity to data partitioning and indicates mild overfitting. This effect was expected since a small dataset restricts the intra-class variability available during training. A practical direction to mitigate this issue is the incorporation of audio data augmentation techniques, such as time-stretching, pitch-shifting and additive environmental noise. By the help of confusion matrices shown by Table 11, it can easily be inferred that the ForestX-Net correctly classified 137 out of 148 test images. Only 11 images were misclassified by the proposed model. 15 test images belong to each of three classes were classified without missing. These classes are Fire, Thunderstorm and Chainsaw. The ForestX-Net is remarkable over detecting illegal activities such as damaging the natural life having the 100% accuracy in detecting the "chainsaw" sounds in forest proveing that the proposed model can be used for detecting illegal activities.

6. Conclusion & future directions

In this work, the ForestX-Net, a cross deep neural network, was proposed as a novel solution to the forest sound recognition task. The ForestX-Net outperformed many state-of-the-arts in terms of classification test accuracy. Future studies could investigate data augmentation techniques or transfer learning with pre-trained models to enhance performance further, especially in low-data scenarios. It was planned to integrate data augmentation techniques (as mentioned in Discussion section) in future iterations of ForestX-Net. The model may also be deployed on a device (a kind of artificial ear, an early warning system), and can be put anywhere in a forest to detect illegal activities and forest fire when the incident starts just before it's too late to take an action by official institutions. Also, the ForestX-Net can constitute an inspiration to the researchers by being a starting point of working in forest sound recognition problem.

Declaration of ethical code

The author of this article declares that the materials and methods used in this study do not require ethics committee approval and/or legal-special permission.

Conflicts of interest

The author declares that there is no conflict of interest.

References

Abdoli, S. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, *136*, 252–263. https://doi.org/10.1016/j.eswa.2019.06.016

Abraham, K., Kumar, A., Krishna, A., & Jha, S. (2023). Classification and detection of natural disasters using machine learning and deep learning techniques: A review. *Earth Science Informatics*, *17*, 869–891. https://doi.org/10.1007/s12145-023-01062-3

Akbal, E., Doğan, S., & Tuncer, T. (2022). An automated multispecies bioacoustics sound classification method based on a nonlinear pattern: Twine-pat. *Ecological Informatics*, *68*, 101529. https://doi.org/10.1016/j.ecoinf.2022.101529

Arafath, K. M. I. Y., & Routray, A. (2025). Detection of breath sounds in speech: A deep learning approach. *Engineering Applications of Artificial Intelligence*, *141*, 109808. https://doi.org/10.1016/j.engappai.2024.109808

Aryal, N., & Lee, S. W. (2020). Attention-based ResNet-18 model for acoustic scene classification. *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge Technical Report*. Retrieved from http://dcase.community

Aslam, M. A., Shaikh, A., Mehmood, A., & Cao, Y. (2024). Underwater sound classification using learning-based methods: A review. *Expert Systems with Applications*, *225*, 124498. https://doi.org/10.1016/j.eswa.2023.124498

Atmaja, B. T., & Akagi, M. (2020). Deep multilayer perceptrons for dimensional speech emotion recognition. *arXiv preprint arXiv:2004.02355*. https://doi.org/10.48550/arXiv.2004.02355

Ayankoso, S., Emmanuel, I., Ponnle, A., Adegboye, M., & Adedokun, O. (2024). Development of long-range, low-powered and smart IoT device for detecting illegal logging in forests. *Journal of Dynamics, Monitoring and Diagnostics*, *3(3)*, 20–28. https://doi.org/10.58979/jdmd.v3i3.150

Aytar, Y., Vondrick, C., & Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. *arXiv preprint arXiv:1610.09001*. https://doi.org/10.48550/arXiv.1610.09001

Bandara, M., Jayasundara, R., Ariyarathne, I., Meedeniya, D., & Perera, C. (2023). Forest sound classification dataset: FSC22. *Sensors*, *23(4)*, 1977. https://doi.org/10.3390/s23041977

Chang, J. W., Ma, H. S., & Hu, Z. Y. (2025). Multi-level transfer learning using incremental granularities for environmental sound classification and detection. *Applied Soft Computing*, *169*, 112619. https://doi.org/10.1016/j.asoc.2024.112619

Chen, Y., Zhao, Y., Qian, Y., Li, K., & Zhang, H. (2022). Effective audio classification network based on paired inverse pyramid structure and dense MLP block. *arXiv preprint arXiv:2211.02940*. https://doi.org/10.48550/arXiv.2211.02940

Costantini, G., Ciccarelli, G., Langiulli, R., Delli Carri, T., Boccignone, G., & Squartini, S. (2022). Deep learning and machine learning-based voice analysis for the detection of COVID-19: A proposal and comparison of architectures. *Knowledge-Based Systems*, *253*, 109539. https://doi.org/10.1016/j.knosys.2022.109539

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, *19(90)*, 297–301. https://doi.org/10.1090/S0025-5718-1965-0178586-1

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Dissanayake, T., Dias, D., Fernando, T., Zhang, L., Bandara, R., & McDonald-Maier, K. (2023). Multi-stage stacked temporal convolution neural networks (MS-S-TCNs) for biosignal segmentation and anomaly localization. *Pattern Recognition, 139*, 109440. https://doi.org/10.1016/j.patcog.2023.109440

Fan, X., Jiang, Y., Zhang, H., Zhang, W., & Lu, X. (2024). A dual adaptive semi-supervised attentional residual network framework for urban sound classification. *Advanced Engineering Informatics, 62*, 102761. https://doi.org/10.1016/j.aei.2024.102761

Fava, A., Bernardi, M. M., dos Santos, L., & Romano, R. (2024). Pre-processing techniques to enhance the classification of lung sounds based on deep learning. *Biomedical Signal Processing and Control, 92*, 106009. https://doi.org/10.1016/j.bspc.2023.106009

Gao, H., Yang, Z., Li, Y., Wang, H., & Zhang, J. (2025). An integrated feature extraction framework of linear multi-layer perceptron to reduce computation complexity for remaining useful life prediction. *Engineering Applications of Artificial Intelligence, 141*, 109846. https://doi.org/10.1016/j.engappai.2024.109846

Goulão, M., Gomes, D., Silva, D. F., Pimentel, D., & Martins, H. (2024). Training environmental sound classification models for real-world deployment in edge devices. *Discover Applied Sciences, 6*, 166. https://doi.org/10.1007/s42452-024-06159-8

Guan, J., Liu, B., Liu, J., Liu, H., & Wu, D. (2023). Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining. In *ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10095820

Han, X., & Peng, J. (2024). Bird sound detection based on sub-band features and the perceptron model. *Applied Acoustics, 217*, 109833. https://doi.org/10.1016/j.apacoust.2023.109833

Hassan, E., Al-Sabaawi, A., Ibrahim, R. W., & Al-Mistarihi, M. F. (2024). Optimizing poultry audio signal classification with deep learning and burn layer fusion. *Journal of Big Data, 11*, 135. https://doi.org/10.1186/s40537-024-00908-0

Haupert, S., Sèbe, F., & Sueur, J. (2022). Physics-based model to predict the acoustic detection distance of terrestrial autonomous recording units over the diel cycle and across seasons: Insights from an Alpine and a Neotropical forest. *arXiv preprint arXiv:2211.16077*. https://doi.org/10.48550/arXiv.2211.16077

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135). IEEE. https://doi.org/10.1109/ICASSP.2017.7952132

Huang, D. M., Zhang, Z., Xu, L., Sun, J., & Xu, W. (2023). Deep learning-based lung sound analysis for intelligent stethoscope. *Military Medical Research, 10*, 44. https://doi.org/10.1186/s40779-023-00479-5

Javaheri, B. (2021). Speech & song emotion recognition using multilayer perceptron and support vector machine. *arXiv preprint arXiv:2105.09406*. https://doi.org/10.48550/arXiv.2105.09406

Khishe, M., Mosavi, M. R., & Samadi, S. (2018). Sim chaotic fractal walk trainer for sonar data set classification using multi-layer perceptron neural network and its hardware implementation. *Applied Acoustics, 137*, 121–139. https://doi.org/10.1016/j.apacoust.2018.03.026

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Krishina, T. B., & Kokil, P. (2023). Automated classification of common maternal fetal ultrasound planes using multi-layer perceptron with deep feature integration. *Biomedical Signal Processing and Control, 86*, 105283. https://doi.org/10.1016/j.bspc.2023.105283

Manivannan, S. (2022). An ensemble-based deep semi-supervised learning for the classification of wafer bin maps defect patterns. *Computers & Industrial Engineering, 172*, 108614. https://doi.org/10.1016/j.cie.2022.108614

Mehrish, A., Kumar, S., & Singh, P. (2023). A review of deep learning techniques for speech processing. *Information Fusion, 99*, 101869. https://doi.org/10.1016/j.inffus.2023.101869

Mushtaq, Z., Qamar, S., & Lee, H. (2021). Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Applied Acoustics, 172*, 107581. https://doi.org/10.1016/j.apacoust.2020.107581

Nanni, L., Brahnam, S., Lumini, A., & Gatta, C. (2020). Animal sound classification using dissimilarity spaces. *Applied Sciences, 6(23)*, 762. https://doi.org/10.3390/app6230762

Nasiri, A., & Hu, J. (2021). SoundCLR: Contrastive learning of representations for improved environmental sound classification. *arXiv preprint arXiv:2103.01929*. https://doi.org/10.48550/arXiv.2103.01929

Nogueira, A. F. R., Silva, F. F., dos Santos, R. R., & Carvalho, A. C. (2022). Sound classification and processing of urban environments: A systematic literature review. *Sensors, 22(22)*, 8642. https://doi.org/10.3390/s22228642

Orosoo, M., Zhang, Y., & Li, X. (2025). Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education. *Alexandria Engineering Journal, 111*, 21–32. https://doi.org/10.1016/j.aej.2024.09.002

Panimalar, S. A., Kumar, S., & Raj, R. (2025). Intensified customer churn prediction: Connectivity with weighted multi-layer perceptron and enhanced multipath back propagation. *Expert Systems with Applications, 265*, 125993. https://doi.org/10.1016/j.eswa.2024.125993

Paranayapa, T., Ranasinghe, P., Ranmal, D., Meedeniya, D., & Perera, C. (2024). A comparative study of preprocessing and model compression techniques in deep learning for forest sound classification. *Sensors, 24(4)*, 2177. https://doi.org/10.3390/s24042177

Peng, L., Wang, H., & Li, Z. (2023). BSN-ESC: A big–small network-based environmental sound classification method for AIoT applications. *Sensors, 23(15)*, 6983. https://doi.org/10.3390/s23156983

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE. https://doi.org/10.1109/MLSP.2015.7324337

Saravanan, K., Vijay, K., & Selvan, R. (2018). How to prevent maritime border collision for fisheries? A design of real-time automatic identification system. *Earth Science Informatics, 12*, 241–252. https://doi.org/10.1007/s12145-018-0336-9

Segarceanu, S., Popescu, D., & Andrei, V. (2020). Forest monitoring using forest sound identification. In *2020 IEEE 43rd International Conference on Telecommunications and Signal Processing (TSP)* (pp. 346–349). IEEE. https://doi.org/10.1109/TSP49704.2020.9243641

Shanthakumar, S., Anandan, R., & Kumar, P. (2020). Environmental sound classification using deep learning. *Instrumentation, 7(3)*, 175–183. https://doi.org/10.3390/instruments7030017

Sharma, R., Singh, P., & Kumar, A. (2022). Motor imagery classification in brain-machine interface with machine learning algorithms: Classical approach to multi-layer perceptron model. *Biomedical Signal Processing and Control, 71*, 103101. https://doi.org/10.1016/j.bspc.2021.103101

Simonović, M., Kovandžić, M., Ćirić, I., & Nikolić, C. (2021). Acoustic recognition of noise-like environmental sounds by using artificial neural network. *Expert Systems with Applications, 184*, 115484. https://doi.org/10.1016/j.eswa.2021.115484

Sun, Y., Li, X., Wang, H., & Zhang, Y. (2021). Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks. *arXiv preprint arXiv:2111.14971*. https://doi.org/10.48550/arXiv.2111.14971

Sun, Z., Tao, H., & Li, W. (2024). Broiler health monitoring technology based on sound features and random forest. *Engineering Applications of Artificial Intelligence, 135*, 108849. https://doi.org/10.1016/j.engappai.2024.108849

Tang, C., & Hu, G. (2024). DSCANet: Underwater acoustic target classification using the depthwise separable convolutional attention module. *Earth Science Informatics, 17*, 6123–6135. https://doi.org/10.1007/s12145-024-02538-2

Tripathi, A. M., & Mishra, A. (2021). Environment sound classification using an attention-based residual neural network. *Neurocomputing, 460*, 409–423. https://doi.org/10.1016/j.neucom.2021.06.034

Tripathi, A. M., & Paul, K. (2022). Data augmentation guided knowledge distillation for environmental sound classification. *Neurocomputing, 489*, 59–77. https://doi.org/10.1016/j.neucom.2022.03.011

Tsalera, E., Giannakopoulos, T., & Pikrakis, A. (2021). Comparison of pre-trained CNNs for audio classification using transfer learning. *Journal of Sensor and Actuator Networks, 10(4)*, 65. https://doi.org/10.3390/jsan10040065

Wang, R., Li, X., Zhang, H., & Liu, Y. (2024). A sound event detection support system for smart home based on "two-to-one" teacher–student learning. *Applied Soft Computing, 167*, 112224. https://doi.org/10.1016/j.asoc.2024.112224

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*, 9. https://doi.org/10.1186/s40537-016-0043-6

Wu, S., Zhang, L., Chen, Y., & Li, X. (2024). CRATI: Contrastive representation-based multimodal sound event localization and detection. *Knowledge-Based Systems, 305*, 112692. https://doi.org/10.1016/j.knosys.2024.112692

Xiang, M., Liu, H., Chen, J., & Zhang, Q. (2023). Research of heart sound classification using two-dimensional features. *Biomedical Signal Processing and Control, 79*, 104190. https://doi.org/10.1016/j.bspc.2022.104190

Xiao, H., Li, F., & Wang, Y. (2022). AMResNet: An automatic recognition model of bird sounds in real environment. *Applied Acoustics, 201*, 109121. https://doi.org/10.1016/j.apacoust.2022.109121

Xu, S., & Chen, Y. (2024). Sound classification with time-frequency features in forest environment. *Journal of Physics: Conference Series, 2756*, 012001. https://doi.org/10.1088/1742-6596/2756/1/012001

Yeh, W. C., Lin, C. H., & Tsai, Y. H. (2023). Simplified swarm optimization for hyperparameters of convolutional neural networks. *Computers & Industrial Engineering, 177*, 109076. https://doi.org/10.1016/j.cie.2023.109076

Yi, J., Zhang, X., & Li, Y. (2024). SceneFake: An initial dataset and benchmarks for scene fake audio detection. *Pattern Recognition, 152*, 110468. https://doi.org/10.1016/j.patcog.2024.110468

Yurdakul, M., & Tasdemir, S. (2023). Acoustic signal analysis with deep neural network for detecting fault diagnosis in industrial machines. *arXiv preprint arXiv:2312.01062*. https://doi.org/10.48550/arXiv.2312.01062

Zaman, K., Ahmed, S., & Khan, M. (2023). A survey of audio classification using deep learning. *IEEE Access, 11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3245678

Zhang, H., Li, Y., Wang, J., & Zhao, X. (2025). Sequence–spectrogram fusion network for wind turbine diagnosis through few-shot time-series classification. *Advanced Engineering Informatics, 64*, 102976. https://doi.org/10.1016/j.aei.2024.102976

Zhao, Y., Wang, H., & Li, F. (2022). Deep learning classification by ResNet-18 based on the real spectral dataset from multispectral remote sensing images. *Remote Sensing, 14(19)*, 4721. https://doi.org/10.3390/rs14194721

Zhiqing, W., Li, H., & Zhang, Y. (2024). Enhancing surgical decision-making in NEC with ResNet18: A deep learning approach to predict the need for surgery through x-ray image analysis. *Frontiers in Pediatrics, 12*, 1023145. https://doi.org/10.3389/fped.2024.1023145

Zhu, H., Li, J., & Wang, S. (2021). A spatial-channel progressive fusion ResNet for remote sensing classification. *Information Fusion, 70*, 72–87. https://doi.org/10.1016/j.inffus.2021.01.007