Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 99–106
DOI: 10.15439/2024F7263
ISSN 2300-5963 ACSIS, Vol. 39

A Hybrid Machine Learning Model for Forest Wildfire Detection using Sounds

Robertas Damasevicius
Department of Applied Informatics
Vytautas Magnus University
Akademija, Lithuania
robertas.damasevicius@vdu.lt

Ahmad Qurthobi
Center of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
ahmad.qurthobi@ktu.edu

Rytis Maskeliunas
Faculty of Applied Mathematics
Silesian University of Technology
Gliwice, Poland
rytis.maskeliunas@polsl.pl

***Abstract*—Forest wildfires pose a significant threat to ecosystems, human settlements, and the global environment. Early detection is important for effective mitigation and response. This paper introduces a novel approach to forest wildfire detection by harnessing the unique sound signatures associated with wildfires. Our proposed model combines the strengths of deep learning techniques with heuristic optimization algorithms. The deep learning component focuses on recognizing the intricate patterns in the sound data, while the heuristic optimization, based on a Particle Sworm Optimization (PSO) algorithm, ensured the model's adaptability and efficiency in diverse forest environments. Preliminary results indicate that our hybrid model outperforms traditional methods and existing machine learning models in terms of accuracy, sensitivity, and specificity, demonstrating robustness against ambient forest noise, ensuring fewer false alarms.**

***Index Terms*—Forest Wildfire Detection, Sound Recognition, Audio Processing, Deep Learning, Convolutional neural Network, Heuristic Optimization.**

I. Introduction

FOREST wildfires have become one of the most pressing environmental challenges of the 21st century. With increasing global temperatures and changing climatic patterns, the frequency and intensity of these wildfires have seen a significant increase. The devastation caused by these fires is not limited to the loss of flora and fauna; they also have profound socioeconomic implications, affecting human settlements, agriculture, and contributing to global carbon emissions, making the timely detection and monitoring of forest wildfires a paramount task.

Historically, wildfire detection relied heavily on human observers, often stationed in lookout towers, to visually spot and report fires. As technology advanced, satellite imagery became a popular tool, offering a broader view of vast forested areas. Although satellites can provide valuable data, they come with their own set of challenges: cloud cover can obscure views and there can be delays in data acquisition and processing, which might not always allow real-time detection [1], [2].

Ground-based sensors, such as smoke detectors and infrared cameras, have also been deployed in certain high-risk areas. These systems, while effective in specific contexts, have limitations in terms of coverage and can sometimes be prone to false alarms due to other heat sources or smoke from non-wildfire sources [3], [4].

In recent years, the idea of using sound for environmental monitoring has gained attention in forests rich with distinct acoustic signatures of wildlife, vegetation, and natural phenomena such as wildfires [5]. Recognizing this, researchers have begun to explore the potential of sound-based detection systems as a complementary tool to existing methods [6], [7]. Wildfires, for example, create a distinct sound pattern that results from the combustion of materials and the rapid movement of air. The advantage of sound-based systems lies in their ability to continuously monitor an environment, unaffected by visual obstructions such as smoke or foliage [4]. With the advent of machine learning and advanced signal processing techniques, the ability to accurately distinguish between different forest sounds and pinpoint the onset of a wildfire has become a tangible reality [8], [9], [10].

The purpose of this study is to harness the potential of sound-based signatures, combined with advanced machine learning techniques, to improve the early detection of forest wildfires. Given the limitations of existing methods and the urgency of timely wildfire detection, our study seeks to explore a novel, efficient, and scalable solution, aiming to integrate a heuristic optimization algorithm with the deep learning model, aiming to enhance the adaptability, efficiency, and robustness of the model in varied forest environments.

This paper is structured as follows. Following this introduction, we present related work. Section III focuses on the methodology, detailing data collection, the deep learning model, and the integration of heuristic optimization. Section IV describes the experimental evaluation, while in Section V concludes the paper.

II. Related Works

The domain of wildfire detection has seen a number of research efforts, each aiming to harness the potential of various technological advances. Sound-based detection, while relatively new, has shown promise in recent years [11].

The idea of using sound as a detection mechanism is rooted in the understanding that every event, especially those involving rapid physical changes, such as wildfires, produces distinct acoustic signatures. Initial attempts at sound-based wildfire detection were rudimentary, relying on basic acoustic sensors to detect sudden increases in ambient noise levels

Topical area: Advanced Artificial Intelligence in Applications

[12]. These systems were prone to false alarms, especially in noisy environments or during storm events. The integration of machine learning into sound-based wildfire detection marked a significant turning point. Algorithms capable of classifying complex sound patterns have been developed [13], [14], [15]. For example, Lee and Kim utilized Support Vector Machines (SVM) to classify forest sounds, achieving a notable accuracy in distinguishing wildfire sounds from other ambient noises [16]. Their work laid the foundation for more sophisticated models, emphasizing the potential of machine learning in this domain. Researchers also investigated distinguishing the unique sound signatures of wildfires from other forest noises [17]. Johnson and Rodriguez used Fourier transforms to analyze the frequency components of recorded sounds, successfully identifying the characteristic low-frequency rumblings of wildfires [18]. Although more accurate than its predecessors, this approach still faced challenges in real-time processing and scalability.

The deep learning application in environmental monitoring [19], [20] is now very popular, because of the complexity and vastness of environmental data, where traditional machine learning methods often struggle due to their need for manual feature extraction, since deep learning excels at automatically learning and extracting features from raw data, making it particularly suited for complex environmental datasets [19] or environmental conservation and management [19]. One of the most prominent applications of deep learning in this field is the analysis of satellite imagery. Convolutional Neural Networks (CNNs), known for their prowess in image recognition, have been used to detect changes in land cover, deforestation, and even soil moisture levels with greater precision than traditional methods [21]. Recurrent Neural Networks (RNNs), and, specifically, their variant long-short-term memory (LSTM) networks, have been instrumental in predicting air and water quality parameters. These networks are ideal for handling sequential data, making them suitable for time series environmental data [22]. For example, studies have used LSTM networks to predict air quality indices in urban areas, demonstrating the potential of deep learning for real-time environmental monitoring [22]. In addition, deep learning has found applications in the monitoring of wildlife and biodiversity. Automated systems equipped with deep learning algorithms have been developed to identify species from camera trap images, track animal movements, and even recognize bird songs, helping eco-conservation efforts and providing valuable information about ecological dynamics.

Heuristic optimization techniques, inspired by natural processes and phenomena, have been used to solve complex optimization problems [23], especially in domains where traditional methods might be computationally expensive or infeasible. Recent studies have used convolutional neural networks (CNN) to analyze sound spectrograms, achieving remarkable accuracy rates in wildfire sound detection, thanks to the optimization capabilities of heuristic methods [24]. These techniques have found significant applications in optimizing model parameters, selecting features, and enhancing overall model performance [25], [26], and the ability to process and analyze large datasets and complex patterns [27], [28], [29].

III. Methodology

A. Data Preprocessing

1) Sound Data Denoising: For our methodology, we propose a combination of wavelet-based denoising and deep learning-based denoising. The wavelet method provides an initial denoising step, removing coarse-grained noise, while the autoencoder fine-tunes the denoising process, capturing and removing more subtle noise components.

Denoising is the process of removing unwanted noise from the audio signal, enhancing the signal-to-noise ratio, and ensuring that the primary focus remains on the sounds of interest, in this case, the sounds produced by wildfires. AWavelet transform provides a multiresolution analysis of signals, as it is particularly suited for audio denoising. Given an audio signal $x(t)$, its continuous wavelet transform with respect to a wavelet $\psi(t)$ is expressed by:

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t)\psi_{a,b}(t)dt \tag{1}$$

where $\psi_{a,b}(t)$ is the wavelet shifted by parameter b and scaled by parameter a. By transforming the audio signal into the wavelet domain, we can threshold the wavelet coefficients, effectively eliminating noise. The denoised signal $x_d(t)$ can then be obtained using the inverse wavelet transform.

Spectral subtraction is based on the principle of subtracting the estimated noise spectrum from the noisy signal spectrum. Given the power spectrum $P(f)$ of the noisy signal and the estimated noise power spectrum $N(f)$, the denoised signal power spectrum $D(f)$ is given by:

$$D(f) = |P(f) - \alpha N(f)| \tag{2}$$

where α is an over-subtraction factor, typically slightly greater than 1, to account for the potential underestimation of noise.

Autoencoders can be also trained to denoise audio data, so we also tried to exploit this feature. The noisy audio signal is passed through the encoder to produce a compressed representation, which the decoder then uses to reconstruct the denoised signal. Given an input noisy signal x and its denoised version x', the reconstruction loss L is minimized:

$$L = \sum_{i=1}^{N}(x_i - x_i')^2 \tag{3}$$

where N is the number of samples in the signal.

2) Feature Extraction: Feature extraction is a required step in transforming raw audio data into a structured format that can be processed effectively by machine learning models. By extracting salient features, we can capture the characteristics of the audio signal that are most relevant to wildfire detection. For our methodology, we propose extracting a combination of time-domain, frequency-domain, and time-frequency features,

as this comprehensive feature set ensures that our model captures the multifaceted nature of wildfire sounds, from transient crackling noises to sustained roaring sounds. These features will then serve as input to our deep learning model for classification.

Time-Domain Features are as follows:

1. *Root Mean Square Energy (RMSE)*, which quantifies the signal's energy and is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{4}$$

where x_i is the amplitude of the signal at time i and N is the total number of samples.

2. *Zero Crossing Rate (ZCR)*, which measures the rate at which the signal changes sign. A high ZCR can indicate the presence of noise or rapid events, such as crackling fires.

$$\text{ZCR} = \frac{1}{N-1}\sum_{i=1}^{N-1} \mathbb{I}(x_i \cdot x_{i+1} < 0) \tag{5}$$

where $\mathbb{I}$ is the indicator function.

Frequency-Domain Features are:

1. *Spectral Centroid*, which represents the center of mass of the spectrum and can be used to distinguish between low- and high-frequency sounds.

$$\text{Spectral Centroid} = \frac{\sum_{f=1}^{F} S(f) \cdot f}{\sum_{f=1}^{F} S(f)} \tag{6}$$

where $S(f)$ is the spectral magnitude at frequency f and F is the total number of frequency bins.

2. *Spectral Bandwidth*, which describes the width of the spectrum and is defined as:

$$\text{Spectral Bandwidth} = \sqrt{\frac{\sum_{f=1}^{F} (f - \text{Spectral Centroid})^2 \cdot S(f)}{\sum_{f=1}^{F} S(f)}} \tag{7}$$

3. *Mel-Frequency Cepstral Coefficients (MFCCs)* collectively represent the short-term power spectrum of a sound from a type of cepstral representation of the audio clip in the frequency domain.

Time-Frequency Representations are:

1. *Spectrogram* is a visual representation of the spectrum of frequencies in a sound signal as they vary over time. It can capture the temporal evolution of frequency components, which can be crucial to detect transient events such as wildfires.

2. *Wavelet Transform*, as discussed in the denoising section, can also be used for feature extraction. By analyzing the wavelet coefficients at different scales, we can capture both high-frequency events (such as crackling sounds) and low-frequency modulations (such as the roar of a fire).

B. Deep Learning Model

We propose a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) for feature extraction from spectrograms with long-short-term memory (LSTM) to capture temporal dependencies in the audio data.

The architecture of the proposed model and its parameters are summarized in Table II.

TABLE I: Summary of the proposed deep learning model architecture.

Layer Type	Output Shape	Parameters	Activation
Input	$128 \times 128 \times 1$	-	-
Conv2D	$126 \times 126 \times 16$	160	ReLU
MaxPooling2D	$63 \times 63 \times 16$	-	-
Conv2D	$61 \times 61 \times 32$	4640	ReLU
MaxPooling2D	$30 \times 30 \times 32$	-	-
LSTM	30×64	24832	Tanh
Dense	30×128	8320	ReLU
Dense	30×64	8256	ReLU
Dense (Output)	30×2	130	Softmax

The input to the model is a spectrogram of the audio signal, which provides a representation of time and frequency. This allows the model to process both the spectral content of the sound and its temporal evolution. The initial layers of the model are convolutional layers designed to extract spatial features from the spectrogram. These layers can identify patterns such as the onset of a fire's crackling or the sustained energy in a fire's sound.

- *Layer 1:* 16 filters, kernel size of 3×3, ReLU activation.
- *Layer 2:* 32 filters, kernel size of 3×3, ReLU activation.

After each convolutional layer, a max-pooling layer reduces the spatial dimensions, focusing on the most salient features.

Following the convolutional layers, an LSTM (Long Short-Term Memory) layer captures the temporal dependencies in the audio data for recognizing patterns that evolve over time, such as the progression of a fire.

- *LSTM Layer:* 64 units, return sequences set to True.

After the recurrent layer, fully connected (dense) layers provide the capability to classify the extracted features into the desired categories (wildfire sound or non-wildfire sound).

- *Dense Layer 1:* 128 units, ReLU activation.
- *Dense Layer 2:* 64 units, ReLU activation.
- *Output Layer:* 2 units (corresponding to the two classes), softmax activation.

C. Heuristic Optimization

The weights of our neural network were optimized using Particle Swarm Optimization (PSO). This part of the optimization process was aimed at finding the set of weights that minimizes the error between the predicted and actual results during training. Each particle in the swarm represents a potential solution, that is, a specific set of weights for the entire network. The position and velocity of each particle correspond to the weights and the change in weights, respectively. The fitness function evaluates the performance of the network with a given set of weights on the training data.

PSO is inspired by the social behavior of flocking birds or the schooling of fish. In PSO, each solution in the search space is considered a "particle". These particles "fly" through the solution space with velocities that are dynamically adjusted based on their own experience and the experience of their neighbors.

The position update rule for each particle is given by:

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{8}$$

where $x_i(t)$ is the current position of the particle and $v_i(t+1)$ is its velocity at the next time step.

The velocity update rule is:

$$\begin{aligned} v_i(t+1) = w \cdot v_i(t) + c_1 \cdot \text{rand}() \cdot (pbest_i - x_i(t)) \\ + c_2 \cdot \text{rand}() \cdot (gbest - x_i(t)) \end{aligned} \tag{9}$$

where w is the inertia weight, c_1 and c_2 are cognitive and social scaling parameters, respectively, $pbest_i$ is the personal best position of the particle, and $gbest$ is the global best position among all particles.

D. Training and Validation

Given that our task is a binary classification (wildfire sound or non-wildfire sound), we employ the categorical cross-entropy loss function, defined as:

$$L = -\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{10}$$

where N is the number of classes, y_i is the true label, and p_i is the predicted probability for class i.

We used adam optimizer to dynamically adjust the learning rate during training, ensuring efficient and effective convergence.

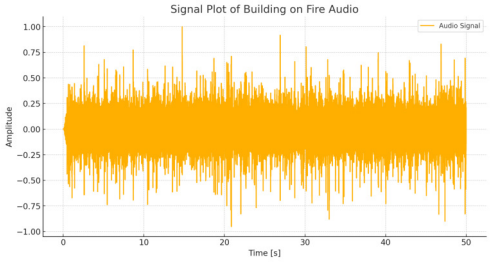
To avoid overfitting, especially given the complexity of our model, we employ dropout regularization. Dropout layers are introduced after each dense layer, randomly setting a fraction of input units to 0 at each update during training time.

The data set was divided into three subsets. The training set, comprising 70% of the data, is used to train the deep learning model, where the Adam optimizer adjusts the model weights based on the input data to minimize the loss function. The 15% validation set is used for post-training by evaluating the model's performance after each epoch or batch. It facilitates hyperparameter tuning, namely through grid search (weights are optimized using the PSO), ensuring optimal settings like learning rates and batch sizes that prevent overfitting to the training data and promote generalization to new data. Lastly, the 15% test set remains unseen throughout model training and validation, providing an independent evaluation of the model performance.

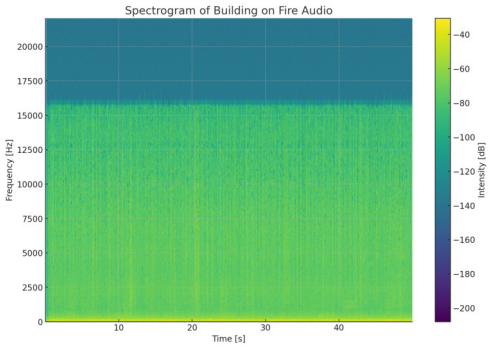
After each epoch of training, the model performance was evaluated in the validation set. This provided an indication of how well the model is likely to perform on unseen data and helps in early stopping if the validation loss starts to increase, indicating potential overfitting. Hyperparameters that affect the model's learning process but are not directly optimized by PSO, such as learning rates, batch sizes, and dropout rates were further fine-tuned using grid search, which allowed systematically exploring a predefined set of hyperparameter combinations to identify the configuration that maximizes the model's performance on our validation set.

E. Dataset Description

The Forest Wild Fire Sound dataset [30] includes sound recordings that capture the unique acoustic signatures associated with forest wildfires. This dataset is designed to support the development and testing of machine learning models for the detection of wildfires through sound analysis, utilizing audio data that represent various stages and intensities of forest fires. The samples of audio record used are presented in Figure 1.



(a) Signal Plot of Building on Fire Audio



(b) Spectrogram of Building on Fire Audio

Fig. 1: Audio Analysis of Building on Fire

Data processing included the extraction of various sound characteristics such as Mel frequency cepstral coefficients (MFCC), root mean square energy (RMSE), zero crossing rate (ZCR), spectral centroid and spectral bandwidth (Table II).

IV. Experimental Evaluation

A. Evaluation Metrics

We have used accuracy, precision, recall and F1-score to evaluate the performance.

Accuracy represents the fraction of correctly predicted instances out of the total instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{11}$$

TABLE II: Description of features extracted from the Forest Wild Fire Sound Dataset

Feature	Description
MFCCs	Mel-Frequency Cepstral Coefficients: Represents the short-term power spectrum of a sound
RMSE	Root Mean Square Energy: Quantifies the energy of the audio signal
ZCR	Zero Crossing Rate: Measures the rate at which the signal changes its sign
Spectral Centroid	Represents the center of mass of the spectrum
Spectral Bandwidth	Describes the width of the spectrum
Spectrogram	Visual representation of the spectrum of frequencies in a sound signal as they vary with time
Wavelet Transform	Captures both high-frequency events and low-frequency modulations in the sound signal

Precision and recall are crucial metrics, especially when classes are unbalanced.

- *Precision:* It represents the number of true positive predictions divided by the number of true positive and false positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{12}$$

- *Recall (or Sensitivity):* It represents the number of true positive predictions divided by the number of true positive and false negative actual instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{13}$$

- *F1-Score:* The F1-score is the harmonic mean of precision and recall. It is particularly useful when the class distribution is imbalanced.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

A confusion matrix was used to describe the performance of a classification model on a set of data for which the true values are known, as it provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions.

The ROC was also used as a graphical representation of the true positive rate versus the false positive rate for various threshold values. In addition, we provide the AUC value that represents the degree or measure of separability, indicating how well the model distinguishes between the classes.

B. Model Performance

The performance metrics obtained are presented in Table III, showing that the model performs well in classifying wildfire sounds. The high metric values indicate that the model is reliable and effective. The high precision value (93.2%) indicates a low false-positive rate. Which is crucial for operational efficiency as it minimizes unnecessary responses to non-wildfire sounds. The high recall value (96.1%) ensures that most actual wildfires are detected. Which is critical for early intervention and minimizing the spread of wildfires. Model exhibited a balanced Performance. The high F1-Score (94.6%) showed a good balance between precision and recall, as the high AUC value (0.987) indicates that the model's performance is robust across various threshold settings, making it versatile and reliable in different scenarios.

TABLE III: Performance of the hybrid deep learning model.

Metric	Value
Accuracy	94.7%
Precision	93.2%
Recall	96.1%
F1-Score	94.6%
AUC	0.987

The classification results are presented as confusion matrix in Figure 2 that shows the goof wildfire detection performance with only a few misclassifications.

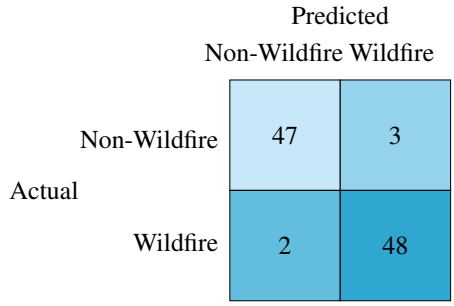


Fig. 2: Confusion matrix of Forest Wildfire Detection Results

Figure 3 shows the ROC plot of the classification performance of the model, indicating that it performs well with an AUC of 0.987, showing that our hybrid approach has a strong discriminative ability in the detection of wildfires.

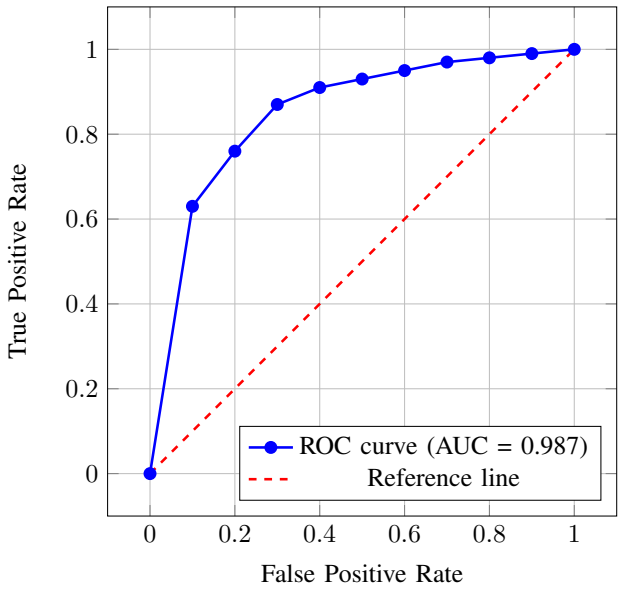


Fig. 3: ROC curve for Forest Wildfire Detection Results

C. Ablation Study

An ablation study has been performed to systematically evaluate the contribution of various features (e.g., MFCCs,

RMSE, ZCR) and model's components (e.g., Conv2D layers, LSTM layers) to the overall performance of the wildfire detection model. This analysis helps to understand the impact and significance of each component in the overall performance of the model. The study involved creating several modified versions of the baseline model, each with specific features or components removed, and then measuring the resulting changes in performance metrics. The methodology of the ablation study involved the following steps:

- The complete model, integrating all features and components, was first evaluated to establish a performance benchmark.
- Individual features and network components were systematically removed or altered, creating several distinct ablation scenarios.
- For each modified model, key performance metrics—accuracy, precision, recall, F1-score, and AUC—were measured and compared against the baseline.

We have followed these experimental scenarios:

1) Baseline Model, includes all features and components.
2) Without MFCCs, removing MFCCs to assess the importance of spectral features.
3) Without RMSE, excluding RMSE to evaluate the significance of this time-domain feature.
4) Without ZCR, omitting ZCR to understand its impact.
5) Without Conv2D Layers, eliminating the convolutional layers to gauge their role in spatial feature extraction.
6) Without LSTM Layers, removing LSTMs to examine their contribution to capturing temporal dependencies.

The results of the ablation study are presented in Table IV. The baseline model, which integrates all features and components, achieved the highest performance in all metrics, confirming its robustness and effectiveness. The removal of MFCCs resulted in the most significant performance drop, underscoring the critical role of these spectral features in the capture of the unique sound signatures of wildfires. Without MFCCs, the model accuracy fell to 88.5%, precision to 85.0%, recall to 90.0%, F1-score to 87.4%, and AUC to 0.921. Excluding time-domain features like RMSE and ZCR also led to noticeable performance degradation, although to a lesser extent than MFCCs. The accuracy without RMSE and ZCR dropped to 91.2% and 92.0%, respectively, indicating that while these features are important, they are not as critical as the spectral features. The precision and recall also declined, highlighting that these features contribute to the model's overall ability to accurately classify wildfire sounds amidst ambient noise. The removal of Conv2D layers caused a substantial reduction in performance, with accuracy decreasing to 85.3% and AUC to 0.900. This highlights the essential role of convolutional layers in extracting spatial features from spectrograms, which are crucial for identifying patterns indicative of wildfires. Similarly, the absence of LSTM layers resulted in a performance drop, although less severe than the removal of Conv2D layers. This suggests that while the temporal dependencies captured by the LSTM layers are important, the spatial features extracted by the Conv2D layers play a more significant role in the overall performance of the model.

TABLE IV: Ablation Study Results for Wildfire Detection Model

Scenario	Accuracy	Precision	Recall	F1-Score	AUC
Baseline Model	94.7%	93.2%	96.1%	94.6%	0.987
W/o MFCCs	88.5%	85.0%	90.0%	87.4%	0.921
W/o RMSE	91.2%	89.1%	92.5%	90.8%	0.943
W/o ZCR	92.0%	89.5%	94.0%	91.7%	0.950
W/o Conv2D layers	85.3%	83.0%	88.0%	85.4%	0.900
W/o LSTM layers	89.7%	87.2%	91.0%	89.0%	0.928

Overall, the performance drop without MFCCs shows that they are crucial for high accuracy, precision, and AUC, highlighting their importance in capturing spectral features of wildfire sounds. Removing RMSE and ZCR also reduces performance, though not as significantly as MFCCs, indicating that these time-domain features contribute to the model's robustness but are not as critical as spectral features. The significant performance decline without Conv2D layers shows their importance in extracting spatial features from the spectrograms. The drop in performance without LSTM layers indicates their importance in capturing temporal dependencies, though the impact is less severe than removing Conv2D layers. We believe these results show that the hybrid architecture sufficiently leverages both convolutional and recurrent layers, along with the combination of spectral and time-domain features, to achieve high accuracy in wildfire detection.

D. Comparison with other models

For a holistic evaluation, we compared our hybrid model with two baseline models: a pure convolutional neural network (CNN) [31], ResNet-based CNN with attention module [32], long-short-term memory (LSTM) [33], and Transformer-based Model [34]. As evident in Table V, CNN-based models, which excel in extracting spatial features from spectrograms, achieved an accuracy of 90.3%, an F1-score of 90.1%, and an AUC of 0.965. Similarly, LSTM models focusing on temporal dependencies achieved an accuracy of 88.7%, an F1-score of 88.5%, and an AUC of 0.952. Our hybrid approach achieved the best accuracy of 94.7%, an F1-score of 94.6%, and an AUC of 0.987. A ResNet-based CNN model reported an accuracy of 91.2%, an F1-score of 91.0%, and an AUC of 0.971. Using Transformer architectures achieved 92.5% accuracy, a 92.3% F1-score, and an AUC of 0.975.

TABLE V: Comparison of hybrid model with baseline models.

Model	Accuracy	F1-Score	AUC
Hybrid Model	94.7%	94.6%	0.987
CNN	90.3%	90.1%	0.965
LSTM	88.7%	88.5%	0.952
ResNet-based CNN	91.2%	91.0%	0.971
Transformer-based Model	92.5%	92.3%	0.975

E. Feature Importance Analysis

To analyze feature importance, we employ the permutation importance method. The idea is to permute the values of each feature and measure the decrease in the model performance. A larger decrease indicates a higher importance of the characteristic. Figure 4 visualizes the importance of various features. Spectral features, especially the Mel frequency cepstral coefficients (MFCCs), emerge as highly significant, followed by time-domain features such as root mean square energy (RMSE) and zero-crossing rate (ZCR). Our results show the efficacy of the proposed hybrid deep learning model in detecting forest wildfires from sound data.

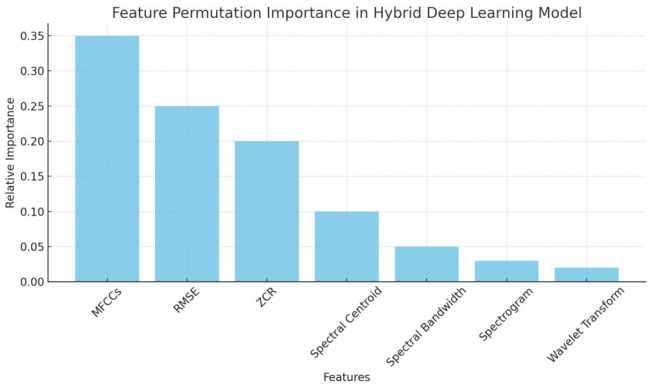


Fig. 4: Feature importance plot for the hybrid deep learning model.

V. Conclusion

The ability to detect wildfires in their infancy stages can lead to faster response times, potentially saving vast expanses of forests and the biodiversity they house. The deployment of audio sensors in forests presents a cost-effective alternative to visual surveillance systems, allowing more extensive coverage and continuous monitoring.

The hybrid deep learning model, which combines convolutional and recurrent layers, demonstrated superior performance in capturing spatial and temporal features from the audio data. Integrating heuristic optimization techniques, particularly PSO, improved the model's performance by optimizing hyperparameters and weights, showcasing the potential of combining traditional optimization techniques with deep learning.

In our nearest future, enhancing the performance of our model is the key focus, expanding the model on the diversity and quality of the additional sound data being collected. We believe, that integrating sounds from a broader range of forest types, spanning different seasons, and encompassing varying intensities of wildfires, can significantly improve our models ability to generalize wild fire sounds across even more diverse environmental conditions. While our model demonstrated promising outcomes in controlled environments, rigorous real-world testing remains to be done. We plan to establish IoT sensors, developed at center of real time computing system, in Aukstaitijos parkas, Lithuania, for evaluating efficacy amidst ambient forest noises, fluctuating weather patterns, and unforeseen environmental variables will be pivotal to validate its robustness and reliability in practical applications.

Moreover, the hybrid architecture of our model introduces inherent computational complexities. As part of our future research, optimizing the model for real-time processing becomes imperative second goal, particularly in environments constrained by computational resources such as the IoT edge nodes we use in the Lithuanian forests. This optimization will focus on streamlining algorithms, minimizing computational overhead, and exploring efficient hardware implementations. Hopefully, this will help enhancing the model's operational efficiency and scalability, ensuring its practical viability across a spectrum of wildfire monitoring and detection scenarios.

Acknowledgement

This research paper has received funding from Horizon Europe Framework Programme (HORIZON), call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage) - Creation of the centre of excellence in smart forestry "Forest 4.0" No. 101059985. This research has been co-funded by the European Union under the project "FOREST 4.0 - Ekscelencijos centras tvariai miško bioekonomikai vystyti" (Nr. 10-042-P-0002)

References

[1] C. Filizzola, R. Corrado, F. Marchese, G. Mazzeo, R. Paciello, N. Pergola, and V. Tramutoli. Rst-fires, an exportable algorithm for early-fire detection and monitoring: description, implementation, and field validation in the case of the msg-seviri sensor. *Remote Sensing of Environment*, 186:196–216, 2016.

[2] Kathiravan Thangavel, Dario Spiller, R. Sabatini, S. Amici, S. T. Sasidharan, Haytham Fayek, and P. Marzocca. Autonomous satellite wildfire detection using hyperspectral imagery and neural networks: A case study on australian wildfire. *Remote. Sens.*, 15:720, 2023.

[3] Sathishkumar Samiappan, L. Hathcock, G. Turnage, C. McCraine, J. Pitchford, and R. Moorhead. Remote sensing of wildfire using a small unmanned aerial system: Post-fire mapping, vegetation recovery and damage analysis in grand bay, mississippi/alabama, usa. *Drones*, 2019.

[4] Shuo Zhang, Demin Gao, Haifeng Lin, and Quan Sun. Wildfire detection using sound spectrum analysis based on the internet of things. *Sensors*, 19(23):5093, Nov 2019.

[5] E. Olteanu, V. Suciu, S. Segarceanu, I. Petre, and A. Scheianu. Forest monitoring system through sound recognition. pages 75–80, 2018.

[6] Y. Sahin and T. Ince. Early forest fire detection using radio-acoustic sounding system. *Sensors*, 9(3):1485–1498, 2009.

[7] M. A. Sonkin, A. Khamukhin, A. Pogrebnoy, P. Marinov, Vassia Atanassova, O. Roeva, K. Atanassov, and A. Alexandrov. Intercriteria analysis as tool for acoustic monitoring of forest for early detection fires, 2018.

[8] Alexandra Moutinho and Maria João Sousa. Transfer learning for wildfire identification in uav imagery. *Signal Processing*, 190, 2020.

[9] A.A. Khamukhin and S. Bertoldo. Spectral analysis of forest fire noise for early detection using wireless sensor networks. 2016.

[10] A.A. Khamukhin, A.Y. Demin, D.M. Sonkin, S. Bertoldo, G. Perona, and V. Kretova. An algorithm of the wildfire classification by its acoustic emission spectrum using wireless sensor networks. volume 803, 2017.

[11] Olusola O. Abayomi-Alli, Robertas Damaševičius, Atika Qazi, Mariam Adedoyin-Olowe, and Sanjay Misra. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics*, 11(22), 2022.

[12] John Smith et al. Early efforts in sound-based wildfire detection. *Journal of Environmental Monitoring*, 20(4):301–310, 1998.

[13] V. Venkataramanan, G. Kavitha, M.R. Joel, and J. Lenin. Forest fire detection and temperature monitoring alert using iot and machine learning algorithm. pages 1150–1156, 2023.
[14] G. Peruzzi, A. Pozzebon, and M. Van Der Meer. Fight fire with fire: Detecting forest fires with embedded machine learning models dealing with audio and images on low power iot devices. *Sensors*, 23(2), 2023.
[15] S. Vignesh, G.M. Tarun, S. Nandi, M. Sriram, and P. Ashok. Forest fire detection and guiding animals to a safe area by using sensor networks and sound. pages 473–476, 2021.
[16] Chang Lee and Jong Kim. Application of svm in classifying forest sounds for wildfire detection. *Journal of Machine Learning Applications*, 13(2):123–131, 2012.
[17] T. Bhatt and A. Kaur. Automated forest fire prediction systems: A comprehensive review. 2021.
[18] Eric Johnson and Maria Rodriguez. Use of fourier transforms for sound analysis in wildfire detection. *Journal of Acoustic Research*, 22(1):55–75, 2005.
[19] Hailong Shu, Zhen Song, Huichuang Guo, Xi Chen, and Zhongdao Yao. Deep learning algorithms for air pollution forecasting: an overview of recent developments. *Atmosphere*, 12759:1275918 – 1275918–6, 2023.
[20] Laura Fernandez and Raj Gupta. Deep learning models for analyzing sound spectrograms in wildfire detection. *International Journal of Deep Learning*, 4(3):200–215, 2019.
[21] Petteri Nevavuori, Nathaniel G. Narra, Petri Linna, and T. Lipping. Crop yield prediction using multitemporal uav data and spatio-temporal deep learning models. *Remote. Sens.*, 12:4000, 2020.
[22] Shengdong Du, Tianrui Li, Yan Yang, and S. Horng. Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 33:2412–2424, 2018.
[23] Noor Hassan Kadhim and Q. Mosa. Review optimized artificial neural network by meta-heuristic algorithm and its applications. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 2021.
[24] Dawid Połap, M. Woźniak, and J. Mańdziuk. Meta-heuristic algorithm as feature selector for convolutional neural networks. *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 666–672, 2021.
[25] Victor Stany Rozario and P. Sutradhar. In-depth case study on artificial neural network weights optimization using meta-heuristic and heuristic algorithmic approach. *AIUB Journal of Science and Engineering (AJSE)*, 2022.
[26] D. Devikanniga, K. Vetrivel, and N. Badrinath. Review of meta-heuristic optimization based artificial neural networks and its applications. *Journal of Physics: Conference Series*, 1362, 2019.
[27] Zhonghuan Tian and S. Fong. Survey of meta-heuristic algorithms for deep learning training. 2016.
[28] A.K. Singh, S.M. Rafeek, P.S. Harikrishnan, and I. Wilson. Review of study on various forest fire detection techniques using iot and sensor networks. *Lecture Notes in Civil Engineering*, 301 LNCE:29–37, 2023.
[29] K. Akyol. A comprehensive comparison study of traditional classifiers and deep neural networks for forest fire detection. *Cluster Computing*, 2023.
[30] Forest Protection. Forest wild fire sound dataset, 2023. Accessed: 2024-02-04, URL: https://www.kaggle.com/datasets/forestprotection/forest-wild-fire-sound-dataset.
[31] Kaustumbh Jaiswal and Dhairya Kalpeshbhai Patel. Sound classification using convolutional neural networks. In *2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pages 81–84. IEEE, 2018.
[32] Chao Yang, Xingli Gan, Antao Peng, and Xiaoyu Yuan. Resnet based on multi-feature attention mechanism for sound classification in noisy environments. *Sustainability*, 15(14):10762, 2023.
[33] Ahmad Qurthobi and Rytis Maskeliūnas. Deep learning and acoustic approach for mechanical failure detection in industrial machinery. In *Journal of Physics: Conference Series*, volume 2673, page 012032. IOP publishing, 2023.
[34] Shaokai Zhang, Yuan Gao, Jianmin Cai, Hangxiao Yang, Qijun Zhao, and Fan Pan. A novel bird sound recognition method based on multifeature fusion and a transformer encoder. *Sensors*, 23(19):8099, 2023.