International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 16(2024) pp. 70-86

Submitted: 1 Jan, 2018; Accepted: 1 Jan, 2018; Publish: 6 May, 2024

Preventing Illegal Deforestation using Acoustic Surveillance

Madhav Rajesh[1], Taarussh Wadhwa[1], Aswani Kumar Cherukuri[1], Firuz Kamalov[2], Annapurna Jonnalagadda [3] and Santosh Ray[4]

[1]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India.
cherukuri@acm.org
[2]Dept of Electrical Engineering, Canadian University Dubai, UAE.
[3]Shool of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
[4] Dept of Information Technology, Khawarizmi International College, Al Ain, UAE

Abstract: **With the rapid increase in deforestation and the subsequent impact on global warming, rainforest protection is the first step to preventing drastic climate change. Audio classification based on audio recognition techniques is promising as they have consistently performed better than humans in urban sound classification. The challenge arises as the research performed on the audio classification of natural sounds such as the rainforest are in their preliminary stage and the shortage of a strongly labelled dataset. This paper proposes a solution to prevent illegal deforestation in rainforests with acoustic surveillance and deep learning. Further, this works to adopt transfer learning on three different models, YAMNet, AlexNet, and ResNet-50, to discover which methodology yields the most practical and effective approach to send real-time alerts for chainsaw incursions in rainforests. We also introduce an architecture that allows our solution to deploy over mobile phones. The investigated method is further extended in an automated prototype that future researchers can easily integrate into solutions based on cloud technology for real-world deployment.**

Keywords: Audio classification, Acoustic surveillance, Computer vision, Convolutional Neural Network

I. Introduction

Rainforest plays an invaluable role in sustaining life. They are the oldest living ecosystem on Earth, with some surviving in their current form for over 70 million years. Although these only accounts for 6% of Earth's surface, they habituate more than half of its animal and plant species. Deforestation is responsible for nearly 20% of all global carbon emissions and accounts for trillions of dollars of economic loss. Rainforests that once grew over 14% of the land on Earth now cover only about 6%. If present deforestation rates continue, these crucial ecosystems may vanish totally from the Earth over the next century [2]. Therefore, protecting rainforests is essential to fighting against climate changes and preserving biodiversity. However, there are several obstacles. The majority of the forest rangers don't have enough resources and workforce to keep an eye on thousands and thousands of acres physically. Along with this, extreme weather conditions also pose a severe challenge. With the assistance of modern technology, tremendous efforts have been undertaken to help conserve rainforests. With the rising advancement of machine learning and artificial intelligence in various businesses and society in general, these sophisticated technologies have piqued the interest of those committed to environmental protection.

This paper builds a solution to protect rainforests with acoustic surveillance and deep learning. We focus on avoiding illegal deforestation and incorporating artificial intelligence to increase the efficiency and efficacy of on-the-ground rainforest protection. Despite recent breakthroughs in various fields, implementing such technology to real-world conservation efforts remains difficult. The deployment of AI-powered devices in rural locations is constrained by several factors, including limited power, inadequate connection, and severe environments [3]. One possible solution is airborne monitoring of the dedicated forests. Ever since the satellite images from the Earth Observation satellites like the Landsat series were made freely available in 2008, scientists have tried to map the tree cover and gather data regarding deforestation patterns [4]. This data is then utilised by the "Brazilian Institute of Environment and Renewable Natural Resources (IBAMA)" to keep an eye on forest cover and send notifications when any significant changes are detected. However, these images are not sharp enough to explain such destruction due to a relatively low resolution. It can be because of wildfires, illegal logging, or even clearcutting. Local environmental agencies must investigate the deforestation warnings to confirm unlawful activities before creating a report, taking around six hours to complete. Even after this, the probability of any sort of action being taken is relatively low. Last year, for example, less than 1% of 150,000 alerts issued across the country resulted in action being taken [11]. MapBiomas is a network of universities, non-governmental organisations, and tech firms that has now invented a technique to detect illegal deforestation in virtually real-time. We created a platform that automatically gathers data from the Brazilian government's existing alarm systems. Then compared it to images acquired by small satellites from Planet Labs, a private business located in San Francisco, CA, with far better-quality images — down to just three metres. However, monitoring deforestation activities across the country using only the data from Planet Labs would be quite expensive and tedious because of the large number of images that would need accurate processing. As a result, the MapBiomas platform decided to only zoom into regions that have been designated as potentially deforested and will then use comprehensive pictures collected before and after the occurrence to generate an automated report for prosecution. These studies suggest that by collecting important data of images of these forests and performing classification tasks on these images, the illegal deforestation in these forests can be curbed. While the discussed methods show specific results, the images generated over tropical forests can be inconsistent due to the frequent overcast conditions that result in an abundance of clouds image data constrained by field of vision and area covered by dense understory. Audio data can be a perfect fit in terms of data collection, durability, small data size, and superior information density.

Despite significant developments in visual deep learning, acoustic deep learning is still largely undeveloped for this problem [8, 9, 10, 12]. We believe there is a huge potential for auditory machine learning and its application in detecting illegal deforestation [34, 35, 36, 37, 38, 39]. Unlike Visual Surveillance, Audio Surveillance is cheaper and, in some cases, more effective in intrusion detection. For a tropical rainforest landscape with little to no light, video surveillance remains both expensive and ineffective[13, 14, 15, 16, 17]. This study describes and introduces a rainforest conservation strategy based on acoustic surveillance and machine learning technology. We aim to determine which methodology yields the most practical and successful solution for sending real-time alerts for chainsaw intrusions in rainforests using transfer learning on three different models: YAMNet, AlexNet, and ResNet-50. To remedy the absence of current datasets and the unavailability of rainforests and illegal loggers in our proximity, we synthesise and augment a rainforest soundscape. We also present an architecture for deploying this system in the forests via mobile phones. Following is the structure of the paper. Section 2 provides the background of the work, observations from the detailed literature analysis, problem definition, and objectives of the current work. The proposed models for the experiment, i.e., YAMNet, AlexNet, and ResNet-50 architecture, are discussed in Section 3. The experimental setup and the investigation are provided in sections 4 and 5, respectively. In section 6, the proposed method is further extended in an automated prototype that future researchers can easily integrate into solutions based on cloud technology for real-world deployment. Deploying the mobile phone solution of the proposed method is discussed in section 7, followed by conclusions.

II. Background

Sound event detection (SED) models based using convolutional neural networks (CNNs) have been promising [5, 6, 7]. We cannot directly apply these proposed models or labelled datasets since there is a huge domain gap between them and our targeted rainforest activities, despite their promising outcomes on public datasets for research. The efficiency of the computation is also an issue. Extensive research has happened in video surveillance during the last few decades [8, 9, 10, 11, 12]. Conventional techniques require hand-crafted features to represent knowledge. Modern deep neural network topologies that depend on convolutional layers have significantly improved this situation[5]. Until recently, it was possible to extract decent levels of feature information from photos. However, deep neural networks can now accurately discern patterns in films. Acoustic surveillance has a significant improvement over Video Surveillance in the use case of rainforest protection. Lighting is a severe limitation to that of Video Surveillance which the dark, dense canopy of the forest hinders and further limits in the nighttime. Mapping large forest areas would require difficulties in strategically positioning cameras without any blind spots. Further, Video Processing is computationally expensive, leading to difficulties in deploying edge devices with lower computational powers for surveillance. Several research in various fields has been undertaken on audio-based multimedia indexing and information retrieval based on semantic input. Some studies [13, 14, 15, 16, 17], for example, describe strategies for searching an audio file based on the description provided in each of its classes. However, they can be called incomplete solutions for this study because this research aims to translate the semantic description of the classes into a conception that can be easily used to match the retrieved features of audio recordings of any length.

A. Literature Survey

Hershey et al. [5] have used multiple convolutional network architectures to classify multiple audio clips of a dataset with about 70million training set samples, 5240 thousand hours, including 30,871 labels that are as good as ones found in videos. The authors carefully studied fully connected deep neural networks, including the "AlexNet", "VGG", and "ResNet" architectures. They also experiment with changing the size of both the training sets and the labels, realising that analogs of the convolutional networks used in classification of images perform well on the particular sound classification tasks, and training and label sets of larger size don't make a significant impact. The authors also found that a model that uses embeddings from previously trained classifiers can perform much more efficiently than raw features on the "Acoustic Event Detection" classification task. The results show that well trained and deep image networks can achieve accurate results on classifying audio samples compared to a simple fully connected networks or pre-trained image classification networks. Results also indicated that training the model on a bigger label set may improve the overall performance while evaluating smaller label sets. The paper also found that regularisation may reduce the gap between the models that have been trained on smaller sized datasets as well as larger datasets.

Piczak [6] has evaluated the ability of CNNs to classify audio clips of shorter duration of sound samples from the environmental. A deeply layered model with two convolutional layers along with fully connected layers and max-pooling layers is trained on audio data that are not represented very highly, mainly segmented spectrograms, along with the respective deltas. The accuracy of this model is assessed on three different public datasets of recordings from the environment. The model is able to perform significantly better than some baseline implementations that rely on the ”Mel-frequency cepstral coefficients”. It also achieves results akin to other advanced methodologies. Experiments that were further conducted indicate that a convolutional model achieves a similar level as other feature learning methods and performs better than ordinary approaches based on manually engineered features. Despite accounting for a longer training time, the result isn’t groundbreaking. It indicates that CNNs can be implemented in audio classification tasks involving environmental sounds even with straightforward data augmentation and a limited dataset.

Piczak’s [18] dataset for environmental sound classification shows us that one of the significant drawbacks of research activities focusing on environmental audio classification and detection tasks is the shortage of open-source datasets suitable for adequate research on these topics. In this study, the author tries to help solve this problem by suggesting a new labelled dataset with 2000 short clips that include labels of a multitude of ordinary sounds and a holistic collection of 250 thousand unlabelled auditory clips taken from publicly available recordings on the ’Freesound’ project. This study also includes a detailed evaluation of human accuracy while trying to classify familiar sounds of the environment. This study also compares it with the performance of some basic classifiers that utilise features derived from ”Mel-frequency cepstral coefficients” and ”zero-crossing rate”. The result showed that the classifier performs better in the case of animal sounds than random forest sounds or a combination of these. Although it is a possible artefact of the data, this can also suggest that a suitable option can be utilising customised models for specific wider groups of sounds.

Sailor et al. [19] have shown the ”Convolutional Restricted Boltzmann Machine”, referred to as ConvRBM, functioning as a model for audio and speech signal. The authors were able to construct a ”ConvRBM” using the concept of noisy rectified linear units that provided appropriate sampling. An unsupervised approach was used to train the ”ConvRBM” architecture which helps represent speech signals of unknown lengths. The weights of the discussed model emulate an acoustic ’filterbank’ of sorts. This auditory filterbank is nonlinear with respect to center frequencies of subband filters, like ”Mel”, ”Bark”, ”ERB”, etc. which are some of the standard filterbanks. The authors use this proposed model to learn the correct features applicable to speech recognition tasks. Experiments performed with the ”WSJ0” and ”TIMIT” databases indicate that ”ConvRBM” features can outperform standard spectral features in both hybrid ”DNN-HMM” and ”GMM-HMM” systems. It was observed that the performance of ”ConvRBM” features improved compared to ”MFCC” with a relative improvement of 7% on ”WSJ0 database” and 5% on ”TIMIT” test set for both the sets using ”GMM-HMM” systems.

Takahashi et al. [7] have proposed new CNN architectures and showed that they allow learning a model for end-to-end audio detection by directly modelling several seconds long signals. This is in contrast to previous works and enables the training of end-to-end audio detection. The proposed architecture is inspired by the success of ”VGGNet” architecture. Hence it uses small, 3 by 3 convolutions and a much greater depth than the previous methods. To prevent over-fitting of the model and leverage the proposed network’s modelling capabilities, they further discuss a novel data augmentation method that helps prevent overfitting and leads to significantly better performance even with limited data. Results from the experiments show that the methods significantly outperform previous pioneering approaches. They also validated the effectiveness of large input fields, deep architectures, and data augmentation one by one.

Yusoff et al. [1] have shown advancement in detecting audio events for intrusion detection systems based on various noises like chainsaw activities, wildlife environmental noise and vehicle noises. This experiment uses two main techniques: feature extraction of ”Linear Predictive Coding” and ”Random Forest classification”. The datasets used for testing and training were retrieved from the ”Wildlife Conservation Society” in Malaysia. Experiments demonstrate that this methodology achieves up to 86% accuracy for indicating an intrusion with the help of accurate audio identification. This study proves to be a significant venture as an initial study for classifying data sets of intruders. Wildlife protection agencies can benefit significantly from this intrusion detection. It helps them maintain security as it does not consume as much power as the current surveillance techniques based on camera trapping.

Gemmeke Jort et al. [20] have described the development of ”Audio Set”, a large-scale, open-source data set of manually labelled audio events that mainly aims to solve the lack of data availability research activities involving Image and audio. The authors use a delicately structured hierarchical ontology of 632 different audio classes based on manual curation and the literature. Through this, they can obtain data from human labels to investigate the presence of specific audio classes in short segments, about 10 seconds In duration, of different YouTube videos. Segments are suggested for labelling based on searches using the available context, metadata, and analysis of the content. This experiment produces a dataset of great depth and breadth that aims to stimulate the development of event detectors and classifiers that can perform well.

Lie Lu et al. [21] have presented research regarding audio content analysis for classifying and segmenting, wherein audio streams are segmented according to the type of audio or identity of a speaker. They suggest a holistic approach that can classify and segment a stream of audio into environment sound, speech, music, and silence. The authors, after adequate research, were able to devise a novel algorithm following the ”linear spectral pairs-vector” quantisation and ”K Nearest Neighbor” approach. Certain new features like the ratio of noise to frame for audio samples and band periodicity have also been introduced as they help improve the accuracy significantly and are discussed in detail. The authors further examine and introduce an unsupervised algorithm to segment the speakers that use a novel methodology based on the fundamentals of ”quasi-GMM” and ”LSP” correlation analysis. With the help of this, they are able to develop an improved unsupervised approach for speaker segmentation.

Kim Taejun et al. [22] have studied ”SampleCNN” and its modified architectures by performing robust experiments with three different datasets of acoustic samples. With the help of a thorough analysis of architectures of the sample-level CNN and comparing them to spectrogram CNN and WaveNet, they were able to verify that the architecture with filters of small size produces the best results. Another exciting find was that striding nearly half the blocks from the bottom can without any significant damage to the performance. Out of all the "SampleCNN" architectures examined, "SE block" is most effective concerning computational efficiency and performance, suggesting that the channel-wise recalibration of feature maps helps avoid or limit the discriminative power in audio detection and classification tasks.

Changsong Yu et al. [23], have proposed a multi-layered attention model that aims to tackle the problem of weakly-labelled audio detection models. The purpose of this study based on the classification of audio is to predict the absence or presence of specific acoustic events in a short clip of an audio event. The authors used the "Audio Set" dataset from Google. A limitation of using this dataset is that clips do not include data such as the onset and offset time of that particular audio event, only the absence or presence of specific acoustic events. The proposed model adds to the previously proposed attention model, which involves a single-level concept. It consists of a series of attention modules that can be applied to intermediate layers in the neural network. The outputs obtained from the attention model layers are collated to a result vector, and to makes the final prediction for each individual class a multi-label classifier is used. Experimental results demonstrate that the model discussed achieves an average precision of 0.360, which performs better than the advanced single-level attention model of 0.327 and Google baseline of 0.314.

Pons Jordi et al. [24] have presented a method to examine architectures of convolutional neural networks by comparing the accuracies of classification tasks obtained while using different convolutional architectures that are randomly weighted as feature extractors. The results obtained are well understood as randomly weighted CNNs are nearly similar to the accuracies of trained CNNs that are able to outperform "MFCC" s. Experimental results show that these architectures prove to be extremely important to deep learning solutions. Subsequently, searching for effective architectures capable of encoding the acoustic signals and their specificities can have a stimulating impact in advancing the state of this domain. To support this, they have also shown that acoustic signatures embedded in structures of the model can capture useful cues for classifying classes related to tempo and rhythm. With the help of this methodology, the authors performed experiments of these proposed architectures for classifying acoustic samples and prove that the architectures are an important aspect for fixing problems that may arise while using deep neural networks for auditory data.

Wyse [25] has shown that some representations and issues that are particularly common related to spectrograms for generating acoustic data using neural network architectures used for style transfer applications. The manner of representing data presented to and subsequently generated by a network is one of the decisions examined while designing a neural network for any application. For tasks involving audio as date, the choice of representing data can be more complicated than it seems to be while working with imagery data. Many representations have been experimented on various use cases, including the hand-crafted features, pure digitised streams, "MFCC" s, features discovered with the help of machines, and variants that include a wide variety of spectral representations. The research shows that spectral representations can have a significant role in certain applications like classification tasks that use neural networks. These representations can represent more information than most customised features traditionally of lower dimension than raw audio and are used for acoustic analysis.

Purwins Hendrik et al. [26] have discussed advanced deep learning approaches for processing acoustic signals used for audio classification, detection, or recognition tasks. Environmental sound, speech, and music processing are considered under the same umbrella. This helps in leveraging the key differences and similarities between the fields, addressing the different issues, essential references, approaches, and possibilities for sharing and using related information between different domains. Feature representations like "raw waveform" and "log-mel spectra" and key deep learning models, such as CNN's and variations of LSTM's, are evaluated. Common deep learning application areas have also been discussed in detail, i.e., audio recognition and production and transformation, like audio enhancement, generative speech and sound synthesis models, and source separation.

Colangelo Federico et al. [27] have proposed architecture for surveillance purposes that are able to recognise suitable audio events. This system depends on a hierarchical classification approach that comprises two recurrent networks capable of detecting whether a significant event is happening or not and subsequently classifying it accurately. In order to evaluate the efficiency of the approach, the authors perform three experiments. The proposed method performed well in all three tests and highlighted the key aspects for improving performance. This hierarchical architecture that has been discussed offers a small yet significant advantage with respect to processing time and accuracy when classifying time segments in the background.

Sacco et al. [28] have shown a novel application for edge computing that can detect human presence in disaster situations and leverages state-of-the-art machine learning approaches. The authors deploy a management architecture to calibrate tasks when the edge networks can be compromised. It is done to guarantee the application's reliability level that is acceptable and speeds up the computation process. The inclusion of this layer takes advantage of the network queues model to approximate all total tasks involved. It allows the application to identify the immediate usage of each IoT device on the network and the average queuing time that each task takes to be executed or transferred to the edge of the network. This management layer proves to be an effective tool for policy programmability of the mission re-planning problem that may arise in any IoT device deployed in compromised environments for the edge network, like rainforests in our case. The time taken for processing audio is fairly reduced when the underlying service is running since the application on top is able to leverage features that are capable of improving the overall performance of the discussed system.

Mporas Iosif et al. [29] presented a framework that uses acoustics to detect illegal deforestation in rainforests. The metric used to evaluate the suggested methodology is the performance of logging detection classification tasks and various other methods and algorithms that are widely used for classification tasks. Multiple experimental setups were followed by adjusting parameters and the "support vector machine approach" reported the best classification accuracy. A post-processing on decision level, which significantly improved accuracy, was also applied. Finally, the authors reviewed a "late-stage fusion method", that combines results of the top-three most accurate classifiers, and results from the experiments showed a further increase in performance of nearly 2%, with audio identification for the logging sound reaching an accuracy of 94.4% for a 20db noise to sound ratio.

Sheikh Fahad Ahmad et al. [30] have talked about deforestation and how forests still cover about 3% of Earth's land. However, due to illegal deforestation, these forests covers are deteriorating at a rate close to half the size of England as each year passes. In most forests, deforestation activities are illegal, but a shortage of human resources and other resources results in authorities rarely detecting unlawful deforestation and curbing this threat. A possible solution to this is to detect the cutting of trees in the initial stages. It ensures that appropriate actions can be taken to curb illegal deforestation. This can be done by monitoring the forest area either manually or with the help of some automatic techniques. The process of cutting trees, no matter how much, usually generates a lot of noise. If we try to leverage this vulnerability and detect this noise by regularly monitoring the acoustic signals in that particular area, significant impact is inevitable. An acoustic signature usually contains valuable information of any such activity in the forest. The authors of this study propose an algorithm for detecting illegal deforestation in forests. The approach discussed is based on K means clustering, "GMM" and principal component analysis for comparison, along with specific distance parameters. The suggested methodology was able to achieve an accuracy of 92

Wrege Peter et al. [31] discussed the accelerating loss of biodiversity across the world, which requires effective methodologies for monitoring and notifying conservation action. In environments such as rainforests and oceans, where direct observation is required, passive acoustic monitoring can collect cost-effective and unbiased data. This method can be applied to a great extent in terrestrial environments, particularly environments like dense tropical rainforests with various noise events occurring simultaneously. The authors are able to show how "PAM" can be used to investigate such behaviour with the help of studies of elephants found in rainforests in Central Africa. The authors also discuss the different approaches and challenges one might face while obtaining such data through acoustics. While such analysis and associated methods are developing rapidly, processing dense raw data requires efficient mechanisms using common hardware and speed advancements in detection algorithms.

Mporas Iosif et al. [32] examine and discuss an approach using acoustic surveillance for automatically identifying unlawful wood logging or tree cutting activities in dense tropical rainforests. The authors examine five different machine learning classification algorithms that use multiple different audio classes to identify chainsaw activity sounds in a turbulent environment like the one found in a rainforest. The authors experimented with different environmental noise interferences, a difference based on the sound-to-noise ratio. Across these different approaches, "Support Vector Machines" delivered the best performance.

Table 1: Summary of literature Survey

Index	**Title**	**Approach**	**Findings**	**Research Gap**
1	CNN architectures for large-scale audio classification[5]	Explore various CNN architectures for acoustic classification task. They also try experimenting with the size of the dataset and labels used.	CNN architectures perform much better on audio datasets than a fully connected neural network designed for specific tasks. Larger datasets and label set slightly improve performance, mostly on smaller datasets.	This research hasn't explored regularisation techniques that may improve performance. Dataset also includes video and not just audio.

Index	Title	Approach	Findings	Research Gap
2	Environmental sound classification with convolutional neural networks[6]	Trained a deep neural network consisting of 2 convolutional layers with 2 fully connected layers and max-pooling with low-level representation of acoustic data. Evaluated the model on three open-source datasets and public recordings too.	The model outperforms Mel-frequency cepstral coefficients-based baseline implementations and provides results comparable to other state-of-the-art techniques.	It is not clear if convolutional models perform satisfactorily with other less complex models, since the focus is only on specific aspects of audio.

Index	Title	Approach	Findings	Research Gap
3	ESC: Dataset for environmental sound classification[18]	Presents a labelled collection of 2000 short clips representing 50 classes of various common acoustic events, available through the Freesound project. It also explores some basic ML algorithms to classify audio samples.	Discussed SVM classifier performs better for sounds of animals than a combination of random forest sounds.	Not an expansive dataset; the number of audio labels is relatively low to be utilised as it is.

Index	**Title**	**Approach**	**Findings**	**Research Gap**
4	Unsupervised Filter-bank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification[19].	ConvRBM using NReLUs is proposed to process puree speech signals.	”ConvRBM features outperform standard spectral features in both ””hybrid DNN-HMM”” and ””GMM-HMM”” systems”	The current model doesn’t model the auditory cortex. The model is not suitable for noise-robust and low resource audio surveillance tasks.
5	Deep convolutional neural networks and data augmentation for acoustic event detection[7]	Proposed a task-specific CNN structure and data augmentation approach for detecting acoustic events that limits over-fitting.	Deeper structures, broad input fields, and data augmentation were all tested individually and shown to be beneficial.	The model cannot yet be used for Acoustic Event Detection in videos or even for summarisation purposes.
6	Acoustic surveillance intrusion detection with linear predictive coding and random forest[1].	Feature extraction of Linear Predictive Coding and Random Forest Classification was employed to develop an intrusion detection based on chainsaw and vehicle engine noises.	Achieved 86% accuracy for detecting an intrusion through audio recognition.	Not applied any deep learning model.
7	Audio set: An ontology and human-labeled dataset for audio events[20].	”This paper entails the development of ””Audio Set””, a labelled data set for Image and audio research purposes.”	Dataset of unprecedented breadth and size that stimulates research of accurate and reliable acoustic event recognisers.	The length of the audio limits usability to an extent, and it isn’t easy to scrape this audio dataset.
8	Content analysis for audio classification and segmentation[21].	The authors provide a research paper on audio content analysis for classification and segmentation, in which an audio stream is segmented based on audio kind or speaker identification.	This model supports multiple speaker modelling, and segmentation in real-time.	Doesn’t assist analysis of video content and indexing. Doesn’t include many audio classes.

Index	Title	Approach	Findings	Research Gap
9	Comparison and analysis of Samplecnn architectures for audio classification[22]	Experimented extensively with sample CNN and its expanded architectures using three different audio datasets.	Successfully explore time-dependent inputs for audio classification tasks in CNN architectures.	The accuracy isn't better than present approaches, for the purpose of audio surveillance.
10	Multi-level attention model for weakly supervised audio classification[23]	Introduced a multi-level attention model in addressing weakly labeled audio classification problems on Audio Set.	Best result of this multi-level attention model successfully exceeds Google's benchmark and the previous state of the art results	Haven't yet combined multi-scale and multi-level features together to train Audio Set.
11	Randomly weighted cnns for (music) audio classification[24]	This paper presents a low-cost method of evaluating CNN architectures by comparing the classification accuracies obtained while employing several randomly weighted CNN architectures as feature extrac-	The results indicate that in order to advance the state of this domain, searching for efficient architectures capable of encoding the specificities of acoustic signals is key.	They haven't explored this methodology on certain aspects and properties of audio, because of which the accuracy isn't compatible with certain existing methodologies.
12	Audio spectrogram representations for processing with convolutional neural networks[25]	This study examines several of these representations as well as the challenges that occur, with an emphasis on spectrograms for audio generation.	Spectral representations can have a crucial role in performance in applications using neural networks for classification or regression.	Other approaches such as exploring frequency bins can provide better performance for audio surveillance tasks.
13	Deep learning for audio signal processing[26]	Provided a good study of deep learning concepts and approaches with respect to Audio Processing.	Audio recognition and synthesis, as well as transformation, are two key deep learning application domains discussed. For audio work, he doesn't stick to one model. Discusses how magnitude spectrograms make reassembling the phase difficult.	Doesn't settle on one model for audio tasks. Clear mention of how magnitude spectrograms pose an obstacle to reconstruct the phase of an audio wave.

Index	Title	Approach	Findings	Research Gap
14	Enhancing audio surveillance with hierarchical recurrent neural networks[27].	This study proposes an algorithm that uses deep recurrent neural network for detecting audio events in chaotic environments.	Improved performances are observed with the help of a hierarchical classification approach composed of two RNNs.	The dataset is relatively small. Doesn't implement CNN as feature extractors.
15	An architecture for adaptive task planning in support of IoT-based machine learning applications for disaster scenarios[28].	This research presents an approach based on deep recurrent neural network for detecting audio events in noisy situations. This study describes a revolutionary edge computing application that can identify the presence of people in catastrophe scenarios using Machine Learning methods.	This approach is successful to reduce the time taken for processing when the underlying service is running.	The approach isn't very feasible for the audio surveillance tasks due to computational costs as well as solution deployment constraints.
16	Illegal Logging Detection Based on Acoustic Surveillance of Forest[29].	The authors present an approach for automatically detecting logging activity in rainforests that makes use of acoustic samples of recordings.	Various commonly used classification approaches and algorithms were examined as part of the evaluation of the framework's logs detection classification performance.	The authors have employed the most advanced accuracy for Support Vector Machines by achieving 94.42% accuracy. However, experimentation in edge devices is yet to be done.
17	Automatic Detection of Tree Cutting in Forests using Acoustic Properties[30].	This work provides an algorithm for detecting logging of trees in forests automatically.	The proposed algorithm uses K-means clustering, GMM and PCA for comparison but is broadly based on the distance between parameters.	Doesn't apply different deep learning techniques on various signal datasets.
18	Acoustic monitoring for conservation in tropical forests: examples from forest elephants. Methods in Ecology and Evolution[31].	The authors demonstrate how PAM may be used to explore communication methods, cryptic behaviour, quantify risks, estimate population size, and assess the success of	The authors compare different strategies against more traditional ones and also discuss the methods, needs, and challenges of procuring acoustic data.	The paper only discusses the different examples to show how useful data on elephant movements can be gathered using terrestrial environments.

Index	Title	Approach	Findings	Research Gap
19	Automatic Forest Wood Logging Identification based on Acoustic Monitoring[32]	The authors of this research developed a technique based on acoustic surveillance for detection of timber cutting activity in forests.	They test five machine learning classification methods for identifying chainsaw wood logging noises in a loud forest setting, employing multiple audio descriptors.	The authors have used Support Vector Machines for classifying and hence only achieved 81.65% accuracy.

III. Deep Learning Models

This study describes and introduces a rainforest conservation strategy based on acoustic surveillance and machine learning technology. We aim to determine which methodology yields the most practical and successful solution for sending real-time alerts for chainsaw intrusions in rainforests using transfer learning on three different models: YAMNet, AlexNet, and ResNet-50.

A. YAMNet

YAMNet (Yet Another Mobile Network) is a deep neural network that uses ”the MobilenetV1(Depthwise-separable CNN) architecture” built by Andrew G et al. (2017)[33] to predict 521 audio event types of the ”Google AudioSet-YouTube corpus”. The implementation of YAMNet begins with one complete convolutional layer, followed by 13 ”depthwise-pointwise” layer pairs with back normalisation and ReLU nonlinear activation function. Strided Convolution is employed for downsampling inputs. The input is then fed into a 1000-wide fully connected softmax classifier following an average pooling. The default implementation of the model generates 521 classes. However, the output has been filtered for the experiment. The array of 521 items is reduced to a two-element array, with each reflecting the event of an intrusion or not. In section 4, the classes are mentioned. The study also assesses the possibility of using Random Forest classifier instead of the softmax classifier. As stated above a 1000-wide fully connected softmax classifier is conventionally utilised at the network's final layer. However, research [36, 37, 38, 39] have been carried out to question this standard. For each model, an instance of Random Forest classifier was trained. Each instance feeds the classifier the output of the given model and produces the identical array of two integers reflecting the event of an intrusion or not. When dealing with multi-row data, the classifier produces an X by 2 size matrix, where X represents the number of rows provided to the model. The architecture of YAMNet is shown below in Figure-1.

B. AlexNet

AlexNet is the CNN architecture that won the the 2012 ”ImageNet Large Scale Visual Recognition Challenge” developed initially by Krizhevsky A et al. (2012)[40]. The Image classifying model consists of eight learned layers. The first 5 are convolutional layers. The first, second and fifth layer also includes a max-pooling operation for feature extraction. The input is then fed into three fully connected layers with dropout. All the layers use a ReLU activation function. A 1000-wide fully connected softmax classifier is conventionally employed for classification. For this study, the original model had been slightly altered. Instead of the normal 1000 classes of the pre-trained model, which has been trained for object detection and image classification at large scale, the last layer now outputs an array of two classes, each reflecting the event of an intrusion or not. We chose to use the model's in-depth features and retrain its eight layers to better match the audio classification problem. The study employs the Random Forest classifier instead of the softmax classifier. For each model, an instance of Random Forest classifier was trained. Each instance feeds the classifier the output of the given model and produces the identical array of two integers reflecting the event of an intrusion or not. The network architecture of AlexNet is depicted below in Figure-2.

C. ResNet-50

Residual Neural Network (ResNet) is a CNN with 50 layers(ResNet-50) that is originally built by He K et al. (2015)[41]. After the success of AlexNet every winning architecture employs more layers for reducing the rate of error. The additional layers help to solve complex computer vision problems more efficiently as the different layers can be trained to achieve efficiency for various tasks leading to higher accuracy. However, this can lead to degradation, which is a phenomenon where the accuracy gets saturated leading to the performance of the model deteriorating. ResNet was employed to solve this issue by ”skip connections” using residual blocks. ResNet-50 consists of 5 stages; each stage includes a convolution block and identity block. Each of the convolution and identity block consists of 3 convolution layers each. The first stage also includes a max-pooling layer with ReLU nonlinear activation function for the purpose of feature extraction. Following an average pooling, the input is then fed into a 1000-wide fully connected softmax classifier. This results in ResNet-50 having over 23 million trainable parameters. For this study, the original model had been slightly changed. The method is similar to that used by AlexNet. Instead of the normal 1000 classes, the last layer now outputs an array of two classes, with each reflecting the event of an intrusion or not. The model's pre-trained version was employed, which is pre-trained to classify images on the MNIST dataset. We choose to employ the model's deep features as well as retrain the model's last nine layers o make it a better fit for audio classification problem. The study employs the use of Random Forest classifier instead of the softmax classifier. For each model an instance of Random Forest classifier was trained. Each instance feeds the classifier the output of the given model and produces the

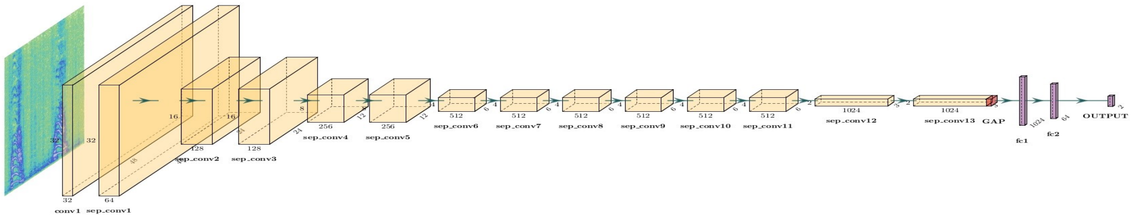


Figure. 1: YAMNet Architecture

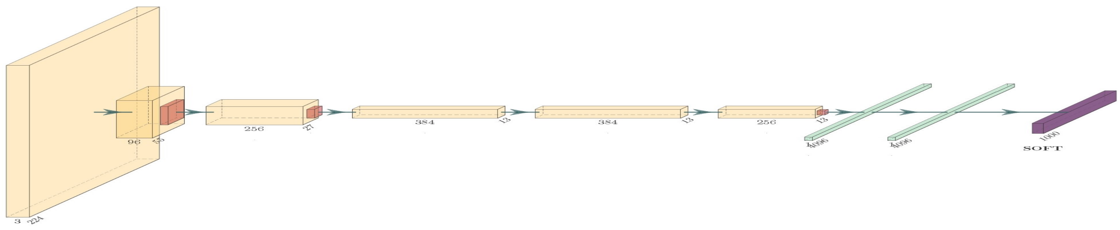


Figure. 2: AlexNet Architecture

identical array of two integers reflecting the event of intrusion or not. is the classifier. The network architecture of ResNet-50 is depicted below in Figure-3.

IV. Experimental setup

This section of the paper presents the audio data synthesised for this experiment. We also illustrate the edge device architecture used to compare each deep learning model's efficiency within the audio classification.

A. Data

The availability of strongly labelled audio data with each event's onset, offset, and semantic description is frequently used to develop novel sound event detection technologies. However, since manually creating such exact annotations takes a long time, few small datasets for sound event recognition have strong labels. Due to the lack of such an existing dataset in literature, and the unavailability of rainforests or illegal loggers to sample audio in our immediate vicinity, we used "Scaper: A library for soundscape synthesis and augmentation" [43] to synthesise and augment a rainforest soundscape. Scaper works as a "high-level sequencer" that can produce different soundscapes from a "single, probabilistically specified specification" given a collection of discrete sound occurrences. The given dataset contains a total of 3240 audio samples of the rainforest. We superimpose various chainsaw noises (from online media sharing platform YouTube and Soundcloud) with different rainforest noises (also from YouTube and Soundcloud), including rain from thunderstorms, wild animals' noises, and water flowing in the rivers of the forest. The samples used for the maximum diversity chainsaw noise class are randomly varied in offset, onset, volume, and pitch. Each of the audio clips is of length 30 seconds and conform to the standard frequency of 44,100 Hz. As seen in Table 2 below, the dataset was divided into a train test split of ratio 80:20.

Table 2: Audio samples collected for each output class

	Rainforest	Chainsaw	Total
Train	2061	800	2861
Test	179	200	379
	2240	1000	3240

B. Edge Device

Our idea of using a mobile phone as an embedded device for acoustic surveillance in the rainforest is based on the capability of mobile phones to record high-fidelity audio from its microphone, which faithfully captures the frequencies produced by chainsaws during illegal deforestation in a rainforest environment. Additionally, mobile phones released in the smartphone era have showcased optimistic processing capabilities to undergo complex computation on the device. Therefore, we propose the usage of such mobile phones for acoustic surveillance under the extreme conditions of rainforest, where the primary challenge is recording faint chainsaw sounds in noise-prone environments. Recycled mobile phones are deployed as sensors but must be enclosed in a waterproof enclosure with solar-powered batteries to support longer battery life and device protection in the rainforest's humid and rainy climate. Figure 4 below shows an illustration of the prototype.

V. Experiment

Deep learning for computer vision applications has shown tremendous progress in the recent past; however, deep learning application for audio research remains in the early stages. Three distinct deep learning models are compared to tackle the challenge of sound classification,. AlexNet, ResNet-50, and YAMNet were chosen for the experiment. The study picked the first two models because they are well-known for classifying images and are extensively used in literature as a baseline for image classification problems. The audio data in this experiment is converted to log-scaled Mel spectrograms; hence image classifiers are a natural fit to the problem [38, 39]. The TensorFlow team has already pre-trained

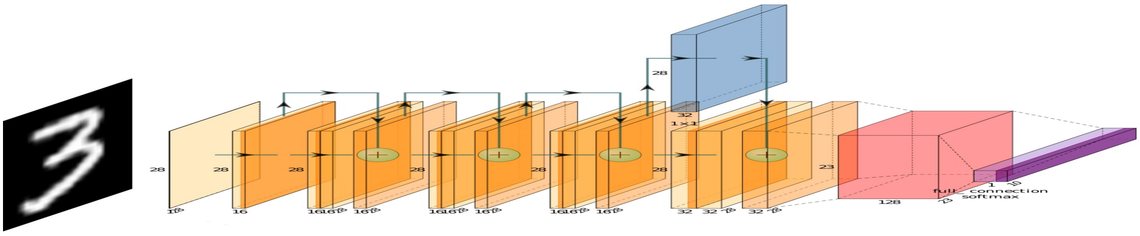


Figure. 3: AlexNet Architecture

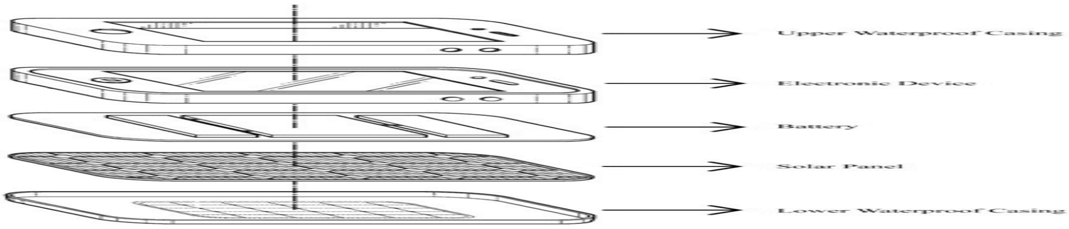


Figure. 4: AlexNet Architecture

the YAMNet model. YAMNet is pre-trained on the same data AlexNet, and ResNet-50 will be trained. YAMNet analyses a sound sample's waveform and predicts the likelihood of each of its pre-trained 521 classes. Figure-5 is an example of a sound waveform.

Since AlexNet and ResNet-50 are image classification algorithms, initially one needs to turn the waveform into log-scaled Mel spectrograms measured using measures of short-time Fourier transform (STFT) with a window size of 25 milliseconds, a window hop of 10 milliseconds, and a periodic Hann window [5]. The study chose these values based on the original implementations of YAMNet [5]. The conversion of raw audio into images was achieved using Python's "librosa: Audio and Music Signal Analysis" [47] library. Figure-6 illustrates an example of a log-scaled mel spectrogram. Next, we trained the models to categorise images; the following procedures have to be followed to utilise them to classify log-scaled mel spectrograms. First, we retrained a few of the model's final layers. For AlexNet, the model's last eight layers were restrained, and we altered some of the layers to classify two classes rather than 1000 (which is its default). Nine of ResNet-50's final layers were retrained and altered to classify two classes instead of the standard 1000. Random Forest, an additional classifier, was introduced as an extra layer to each of the models. Finally, we trained each model's classifier on the same data .

Figure-7 shows a summary of the Proposed event detection system. (a) audio signals are captured in real-time in the rainforest through the microphones of the edge device (in real-world deployment, a rainforest) (b). Further, the audio waveform recorded by the edge device is transformed into log-scaled Mel spectrograms, which can undergo image classification by Convolutional Neural Network (c). The spectrogram is then inputted into the various models (YAMNet, AlexNet, and ResNet-50) trained through transfer learning for feature extraction (d). Finally, a random forest classifier is used to detect the presence of chainsaw interference or not.

The classifier uses minibatch stochastic gradient descent with Nesterov Accelerated Gradient during training to maximise training speed and significantly improve convergence. The classifier is trained using a train/test split of 80:20 ratio.

We used the procedure given in earlier works [42, 44] to calculate Receiver Operating Characteristic-Area Under Curve (ROC-AUC), Matthews Correlation Coefficient (MCC), Accuracy, Precision, and F1 score to evaluate the model's performance and compare them. True and false positives, as well as negatives, are calculated using the metrics. Table 3 explains the difference between true and false positives and negatives.

Table 3: Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN) A	False Positive (FP) B
Actual Positive	False Negative (FN) C	True Positive (TP) D

The receiver operating characteristic (ROC) is a graph that visualises True Positive(TP) against False Positive (FP). We calculate the corresponding True Positive Rate and False Positive Rate for every threshold in a single graph. Higher True Positive rates and lower False Positive rates are desirable. AUC stands for the area Under the Curve. Once the ROC graph is plotted, the AOC of the corresponding graph is calculated. Higher AOC is desirable. The Matthews correlation coefficient (MCC) is most often used in machine learning applications to assess the validity of data element statuses and prediction outcomes (class labels). The metric is often recognised as a more reliable measure, particularly on imbalanced datasets [44].

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

Correctly predicted instances and all instances in the dataset are called accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

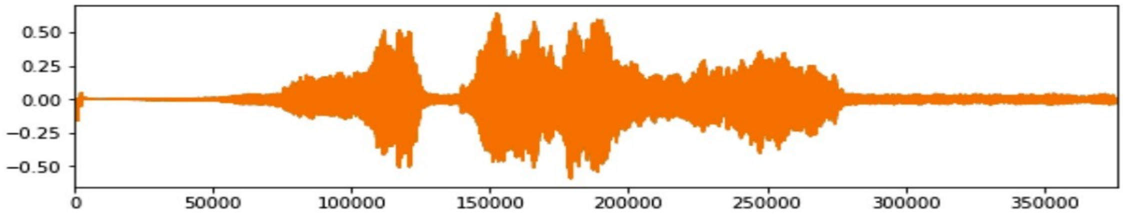


Figure. 5: An example of a sound waveform

(a) Rainforest without chainsaw noise (b) Rainforest with chainsaw noise

Figure. 6: Log-Mel-spectrogram obtained by transforming a sound waveform

Precision refers to the number of positive predictions that were true among the total number of positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

F1 score, also known as F-measure, is the statistical analysis of binary classification that attempts to integrate precision and recall issues in a single metric. The following equation is used to determine the F measure:

$$F_1 = 2\frac{Precision.Recall}{Precision + Recall} \tag{4}$$

The study used the above metrics to provide a complete statistical analysis of the method's effectiveness with each model.

VI. Results and Analysis

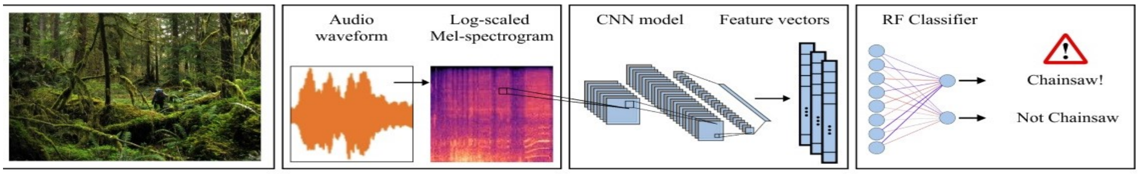
The study tested the models on the evaluation dataset containing 379 audio samples to see how well they performed on fixed-size audio samples. The Performance measure (in %) on classifying fixed-size audio samples for all three models is shown in Table 4. We've also compared two YAMNet output options: pure output (default softmax classifier) from which the classifier extracted two classes and a mix of YAMNet output and Random Forest output (RF). The study did this experiment to understand better whether we should include Random Forest with YAMNet for real-world deployment after the investigation. Figures-**??** and Figures-**??** illustrate the comparison and the ROC curve, respectively.

VII. Deploying on Edge Device with TensorFlow Lite for real-world sounds in rainforests

The models were deployed on an Android device with TensorFlow Lite (TFLite) to compare their performance in real-world scenarios. TFLite [45] is optimised for on-device machine learning. In environments like rainforests, it addresses the challenges concerned with latency and internet connectivity. There is no round trip required to the server, and the trained models can run smoothly without an internet connection. Since they take less RAM space to run on the device, the power consumption is less. Hence, using solar-powered batteries in our device architecture proves beneficial computationally and economically. It is made possible by a key concept called Quantization[46]. In simple words, quantisation uses lower-bit representations instead of a higher-bit representation for a real-valued number. Weights and biases (or simply neural network parameters) are stored as 32-bit single-precision floating-point images scaled between zero and one to perform high-precision calculations during training in deep learning models. Once training is complete, it can be reduced to an unsigned 16-bit integer (2x size reduction) or an 8-bit integer (4x size reduction) representation of the Image, eventually reducing the model size. In many circumstances, it has also been empirically demonstrated that a quantisation leads to limited or no decay, especially when employing 16-bit integer (2x size reduction) representation; hence there is no substantial influence on model correctness. The quantisation is used to reduce the latency and size of the model with a negligible decrease in accuracy. Figure-10 shows the CNN model for mobile applications, and Figure 11 shows the mobile interface of the application.

Table 4: Performance measure (%) on classifying audio samples of fixed size

	YAMNet	**YAMNet with Random Forest Classifier**	**AlexNet with Random Forest Classifier**	**ResNet-50 with Random Forest Classifier**
ROC - AUC	63.5	92	87.5	88.5
MCC	25.1	80.47	64.2	66.4
Accuracy	80.84	91.3	87.3	87.15
Precision	38.5	67.25	60.26	62.89
F1 score	43.15	74.75	70.6	71.27



(a) Audio Capture (b) Post-processing (c) Feature Extraction (d) Event Detection

Figure. 7: Proposed event detection system which goes from capturing a sound to detecting an event of an incursion.

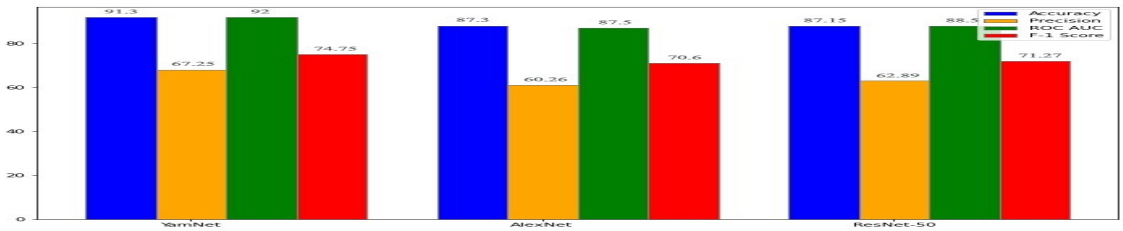


Figure. 8: Comparison of the metrics on the classification problem

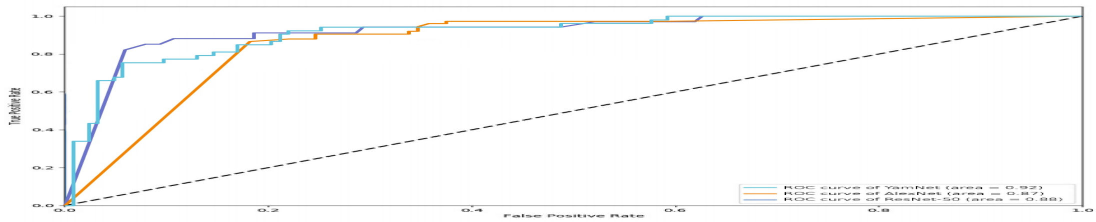


Figure. 9: ROC curves of Chainsaw sound classification in rainforests

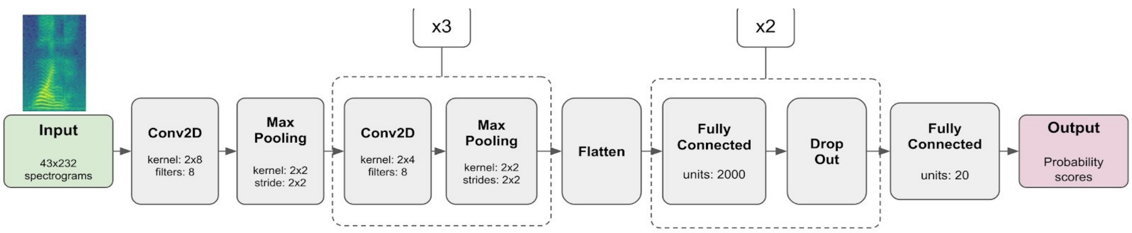


Figure. 10: CNN Model Architecture for Mobile Application using YAMNet

The model that Edge Device uses YAMNet to classify an audio sample of one-second length. The reduced one-second size is done to accommodate the processing capabilities of the edge device. As in the case of YAMNet, the log-scaled mel spectrogram is first processed through a sequence of Conv2D layers with the max-pooling operation for feature extraction. The input is then sent to a sequence of top layers interspersed with dropout. The model's final output is an array of two integers, reflecting the probability scores of the intrusion event.

VIII. Conclusion

This study identifies the detection of illegal logging in rainforests using the Convolutional Neural Network method. The experimental research compared three models' performance, YAMNet, AlexNet, and ResNet-50, classifying rainforest and chainsaw noises. We present a method to synthesise and augment a rainforest soundscape for training the Convolutional Neural Network models. We employed transfer learning to train instances of Random Forest that utilised the model's pre-trained high-level features for all three models (YAMNet, AlexNet, and ResNet-50). The original YAMNet "Google AudioSet-YouTube corpus" dataset includes samples labelled with 521 different classifications. However, we only employed two classes during the experiment to account for the limited processing capability and evaluate the results better. YAMNet correctly identified single fixed-size audio samples 91.3% of the time, AlexNet correctly categorised single fixed-size audio samples 87.3% of the time, and ResNet-50 accurately classified single fixed-size audio samples 87.15% of the time. We can see that AlexNet outperforms ResNet-50 and YAMNet is the best. The motivation behind this was to discover which methodology yields the most practical and effective approach to send real-time alerts for chainsaw incursions in rainforests. Given the limited sample size of the dataset and the unavailability of rainforests and illegal loggers in our immediate vicinity, one immediate direction for future work is working with a bigger and more diverse dataset to accommodate the biases data can create. Recording samples of real rainforest noises and flora and fauna present in it can bring additional benefits for the use case of this study. Optimisation of YAMNet's implementation is another step that future researchers and practitioners can take to reduce feature extraction of the model and reduce the computation cost, yielding higher event detection accuracy. Another area for exploration is few-shot learning to improve event recognition. Adding negatively labelled data can also help to improve classification accuracy. With deployment in a rainforest environment, we see audio surveillance solutions based on cloud architecture as a reality. We hope that the presented experiment and framework significantly contribute as an effective solution in developing applications for monitoring and preserving rainforests.

Declaration Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. No funding was received to assist with the preparation of this manuscript.

Acknowledgments

Please provide acknowledgement only after the conclusion section. The authors acknowledge support from ICCNS, grant number 09-123.

References

[1] M. Yusoff and A. S. Md. Afendi. *Acoustic surveillance intrusion detection with linear predictive coding and random forest* in Soft Computing in Data Science, B. W. Yap, A. H. Mohamed, and M. W. Berry, Eds. Singapore: Springer Singapore, 2019, pp. 72–84.

[2] Environmen2021. *Rain forest threats information and facts* May 2021. [Online]. Available: https://www.nationalgeographic.com/environment/article/rainforest-threats

[3] Y. Liu, Z. Cheng, J. Liu, B. Yassin, Z. Nan, and J. Luo. AI for earth: Rainforest conservation by acoustic surveillance,*arXiv preprint arXiv:1908.07517* 2019.

[4] Nast, C. *How satellite sleuths are helping to save the amazon from destruction*, Aug. 2019. [Online]. Available: https://www.wired.co.uk/article/amazon-rainforest-deforestation-satellite-images

[5] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., CNN architectures for large-scale audio classification, *in 2017 IEEE international conference on acoustics, speech and signal processing (icassp). IEEE*, 2017, pp. 131–135.

[6] K. J. Piczak. Environmental sound classification with convolutional neural networks, *in 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP). IEEE*, 2015, pp. 1–6.

[7] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool. Deep convolutional neural networks and data augmentation for acoustic event detection,*arXiv preprint arXiv:1604.07160* 2016

[8] Kim, K. M., Heo, M. O., Choi, S. H., & Zhang, B. T. "Deepstory: Video story qa by deep embedded memory networks". *arXiv preprint arXiv:1707.00836*.2017.

[9] J. Lei, L. Yu, T. L. Berg, and M. Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval.*Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020.

[10] N. Brindha and P. Visalakshi. Bridging semantic gap between high-level and low-level features in content-based video retrieval using multi-stage esn–svm classifier.*Sadhana*, vol. 42, pp. 1–10, 12 2016.

[11] A. Smeaton, P. Wilkins, M. Worring, O. D. Rooij, T.-S. Chua, and H.-B. Luan. Content-based video retrieval: Three example systems from trecvid. *Wiley Periodicals*, Inc., 2008. 17

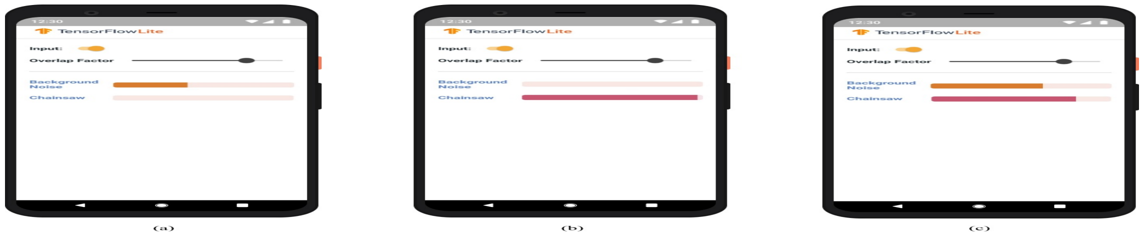


Figure. 11: The mobile application visualisations include an overlay factor that can be manipulated to define the intervals at which samples need to be classified. Image (a) is of audio clips of rainforest, which includes the noise of rain and the fauna of the tropical soundscapes. Image (b) is of only the noise chainsaw, and Image (c) is of an augmented soundscape that includes both the rainforest noise and the chainsaw.

[12] A. Araujo and B. Girod. Large-scale video retrieval using image queries. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1406–1420, 2018.

[13] E. Do Gan, M. Sert, and A. Yazici. A flexible and scalable audio information retrieval system for mixed-type audio signals. *International Journal of Intelligent Systems*, vol. 26, no. 10, pp. 952–970, 2011.

[14] M. Guggenberger. Aurio: Audio processing, analysis and retrieval, in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 705–708.

[15] S. Sundaram and S. Narayanan. Audio retrieval by latent perceptual indexing, in *IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE*, 2008, pp. 49–52.

[16] S.-Y. Kim, K.-M. Kim, and J.-K. Jeon. Quick audio retrieval using multiple feature vectors, in *Digest of Technical Papers International Conference on Consumer Electronics*, 2006, pp. 3–4.

[17] C. Wan and M. Liu. Content-based audio retrieval with relevance feedback, *Pattern Recognition Letters*, vol. 27, pp. 85–92, 01 2006

[18] K. J. Piczak. Esc: Dataset for environmental sound classification, in *Proceedings of the 23rd ACM InternationalConferenceonMultimedia,ser. MM'15*. NewYork,NY,USA:AssociationforComputing Machinery, 2015, p. 1015–1018.

[19] H. B. Sailor, D. Agrawal, and H. Patil, *Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification*, INTERSPEECH, 2017.

[20] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780

[21] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation, *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[22] T. Kim, J. Lee, and J. Nam. Comparison and analysis of samplecnn architectures for audio classifica- tion, *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 285–297, 2019.

[23] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, Multi-level attention model for weakly supervised audio classification, *arXiv preprint arXiv:1803.02353*, 2018.

[24] J. Pons and X. Serra. Randomly weighted cnns for (music) audio classification, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340,2019.

[25] L. Wyse. Audio spectrogram representations for processing with convolutional neural networks,*arXiv preprint arXiv:1706.09559*, 2017.

[26] H.Purwins, B.Li, T.Virtanen, J. Schluter, S.Y. Chang, and T. Sainath. Deep learning for audio signal processing, *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[27] F.Colangelo, F.Battisti, M.Carli, A.Neri, and F. Calabr O. Enhancing audio surveillance with hierarchical recurrent neural networks, in *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2017.

[28] Sacco, Alessio, Flocco, Matteo, Esposito, Flavio, Marchetto, and Guido. An architecture for adaptive task planning in support of IoT-based machine learning applications for disaster scenarios Computer communication,*Computer Communications*, pp.769–778, 2020.

[29] I. Mporas, I. Perikos, V. Kelefouras, and M. Paraskevas. Illegal logging detection based on acoustic surveillance of forest, *Applied Sciences*, vol. 10, no. 20, 2020.

[30] S. F. Ahmad and D. K. Singh. Automatic detection of tree cutting in forests using acoustic properties, *Journal of King Saud University - Computer and Information Sciences*, 2019.

[31] P. H. Wrege, E. D. Rowland, S. Keen, and Y. Shiu. Acoustic monitoring for conservation in tropical forests: examples from forest elephants, *Methods in Ecology and Evolution*, vol. 8, no. 10, pp. 1292–1301, 2017.

[32] I. Mporas and M. Paraskevas. Automatic forest wood logging identification based on acoustic monitoring, in *Proceedings of the 9th Hellenic Conference on Artificial Intelligence, ser. SETN '16*. New York, NY, USA: Association for Computing Machinery, 2016.

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861*, 2017.

[34] J.Wang, P.Huang, Q.Huang, Z.Keand P.Lin. Dialogueactrecognitionforchineseout-of-domainutterances using hybrid cnn-rf, in *International Conference on Asian Language Processing (IALP)*, pp. 14–17, 2016.

[35] L. Zheng, Q. Li, H. Ban, and S. Liu. Speech emotion recognition based on convolution neural network combined with random forest, in *Chinese Control And Decision Conference (CCDC)*, pp. 4143–4147, 2018.

[36] G. Cao, S. Wang, B. Wei, Y. Yin, and G. Yang. A hybrid cnn-rf method for electron microscopy images segmentation, *Journal of Biomimetics, Biomaterials, and Tissue Engineering*, vol. 18, 2013.

[37] Y. Zhu, J. Duan, and T. Wu. Animal fiber imagery classification using a combination of random forest and deep learning methods, *Journal of Engineered Fibers and Fabrics*, vol. 16, p. 155892502110093, 01 2021.

[38] T. M akinen, S. Kiranyaz, J. Raitoharju, and M. Gabbouj. Evolutionary feature synthesis for content- based audio retrieval, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, 02 2013.

[39] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification,*2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 2017.

[40] A. Krizhevsky, I. Sutskever and G. Hinton. ImageNet classification with deep convolutional neural networks, *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[41] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition, *arxiv.org/abs/1512.03385*, 2021.

[42] C. Catal. Performance evaluation metrics for software fault prediction studies, *Acta Polytechnica Hungarica*, vol. 9, no. 4, pp. 193-206, 2012.

[43] J. Salamon, D. MacConnell, M. Cartwright, P. Li and J. P. Bello. Scaper: A library for soundscape synthesis and augmentation, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344-348, 2017.

[44] S. Boughorbel, F. Jarray and M. El-Anbari. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, *PLOS ONE*, vol. 12, no. 6, p. e0177678, 2017.

[45] J. Lee et al., On-Device Neural Net Inference with Mobile GPUs, *arXiv preprint arXiv:1907.01989*, 2019.

[46] R. David et al., TensorFlow Lite Micro: Embedded Machine Learning on TinyML Systems, *Proceedings of Machine Learning and Systems*,vol.3, pp.800–811, 2021.

[47] B. McFee et al., librosa: Audio and Music Signal Analysis in Python, in *Proceedings of the 14th python in science conference*, 18–24, 2015.