Acoustic Environment Identification using Semi-Supervised Learning

MRUDULA JADHAV, ADVAIT DIXIT, AND RASHMIKA PATOLE, *AES Member*,
(jadhavmr18.extc@coep.ac.in) (dixitaa18.extc@coep.ac.in) (rkp.extc@coep.ac.in)

College of Engineering, Pune, India

Audio forensics deals with the authentication and verification of audio recordings that can be used as evidence in the law court and for other litigation purposes. Acoustic Environment Identification (AEI) is a part of Audio forensics that entails identification and verification of the physical environment in which the audio was captured. This work aims to authenticate the audio by identifying the environment it was recorded in using Semi-supervised Learning. We have used a combination of the DCASE2016 and 2017 datasets along with Π-model for Semi-supervised learning. A comparison of supervised and semi-supervised methods has been provided. The results highlight the advantages of using this approach over the general Supervised techniques. This technique generalises the decision boundary better as a result of using a combination of both unlabelled as well as labelled data in the training process.

0 INTRODUCTION

Audio forensics includes the capturing, processing, and interpretation of audio recordings that can be used as evidence in the law court and for other official purposes. It not only involves verification of the direct speaker but also identifying the recording environment, which can be used to determine the underlying facts of an evidentiary recording. Over the years, a great deal of work has been done in the field of Acoustic Environment Identification (AEI). Artifacts embedded inside the audio file during the capturing process can be used to characterize the environment it was recorded in. Two major artifacts used for this purpose are Reverberation and Background noise.

Audio Reverberation is a phenomenon where sound persists even after it has been stopped due to reflections from surfaces such as furniture, people, air, and other objects within a closed surface. Different levels of reverberation time are a result of differences in the geometry and composition of a room. There is a substantial amount of work done in modelling and estimating audio reverberation. Methods in [1], [2] discuss how audio reverberation can be modeled, estimated, and used in a forensic setting. Furthermore, [3] and [4] show that acoustic reverberations can withstand lossy compression and hence can be employed reliably for digital audio forensic applications.

Background noise is also considered as an important acoustic signature for the purpose of environment classification. It depends on the secondary audio source activity and hence the classification decision can be done based on its consistency in the temporal domain. Various noise estimation approaches have also been proposed in [5], [6]. However, the results show that methods using background noise, deliver good performance in some circumstances where background noise is dominant. The system's performance is poor when there is no or very little background noise in the recordings.

Machine Learning (ML) has revolutionized the modern world of computing and has been widely used for multiple applications including AEI. Abundant research has been carried out in the field of ML using different feature sets and models. Features like Mel-frequency cepstral coefficients (MFCC), Temporal Derivative-based Spectrum (TDSM), log Mel spectral coefficients (LMSC) are extracted from the signal. These features are further used for classification using Machine learning algorithms like SVM, logistic Regression, Artificial Neural Networks (ANN).

Few papers [7, 8, 9] present a comparative study of experiments undertaken for the classification of auditory environments using extensive sets of machine learning classifiers along with various acoustic features. However, most of the research is based on supervised learning methods. An approach in [10] uses the Matching Pursuit algorithm on different sets of features for classification but manages to achieve only 65.6% on the DCASE 2016 (Detection and Classification of Acoustic Scenes and Events) dataset. This method is also limited by its high time complexity. There is very little research on unsupervised learning approaches for AEI. Density-based clustering method (DBSCAN) [11] is an unsupervised learning approach used for

AEI but proves to be inefficient since its accuracy and reliability depend on the microphone type used while recording. A large amount of work in machine learning has been done only in the field of Supervised learning. This gives provides a scope to experiment with other techniques like Unsupervised and Semi-supervised learning (SSL). Semi-supervised learning [12], [13] is a method of machine learning in which a little amount of labelled data is combined with a large amount of unlabelled data during training. This paper uses SSL for for acoustic scene classification using the DCASE dataset.

1 METHODOLOGY

The proposed methodology provides a technique to identify the acoustic scene of the audio using Semi-supervised learning. This section is divided as follows:

1) Dataset
2) Data Augmentation
3) Feature Extraction
4) Training

1.1 Dataset

For the task, a combination of DCASE 2016 and 2017 (Detection and Classification of Acoustic Scenes and Events) dataset is employed [14], [15]. The dataset includes recordings from a variety of acoustic environments, each recorded in a distinct location for the same scene. Overall, 15 equally distributed acoustic scenes were used for classification. Figure 1 shows the distribution classes in the data. A 3-5 minute audio was recorded at each location. The original recordings were cut into 10-second segments. There are 546 samples in each audio scenario. The recordings were made using a Soundman OKM II Klassik/studio A3 electret binaural microphone and a Roland Edirol R-09 wave recorder with a sampling rate of 44.1 kHz and a resolution of 24 bits. A total of 8190 (3510 from DCASE 2016 + 4680 from DCASE 2017) audio samples were used for the training purpose.

1.2 Data Augmentation

Data augmentation is a technique for creating synthetic data from an existing dataset to increase the model's generalizability. Usually data augmentation is adopted for increasing the dataset size, however in this case augmentation is performed to introduce perturbations in the data to calculate the mean squared error between the original and the augmented sample. Figure 3a shows the original audio sample in time domain. The different alterations like noise addition, time stretching, pitch shifting and random gain multiplication can be seen in figures 3b, 3c, 3d and 3e respectively.

1.3 Feature Extraction

Extracting features from the audio signal and feeding them to the model yields much better results than feeding the raw audio data into the model. Feature extraction

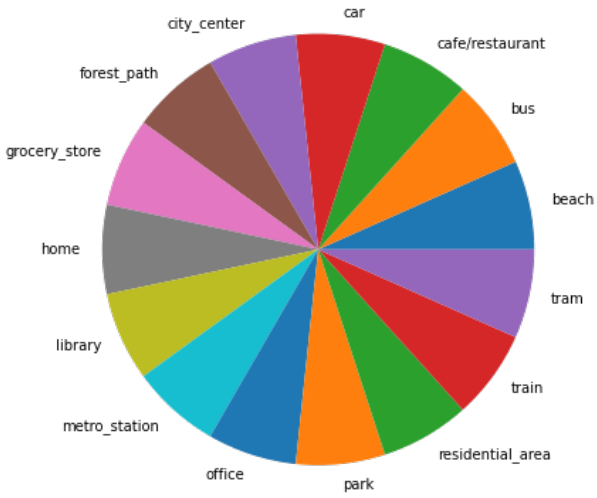


Fig. 1: Distribution of Classes

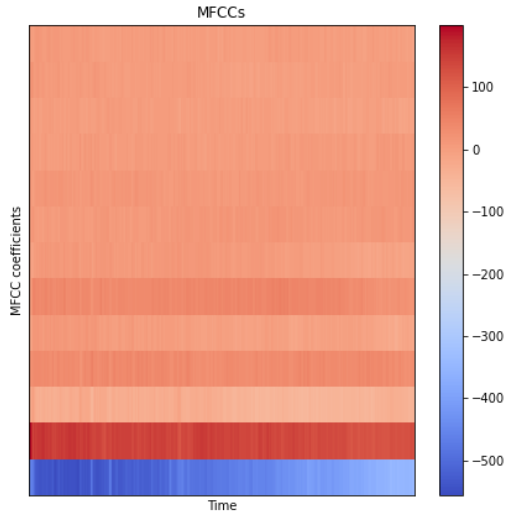
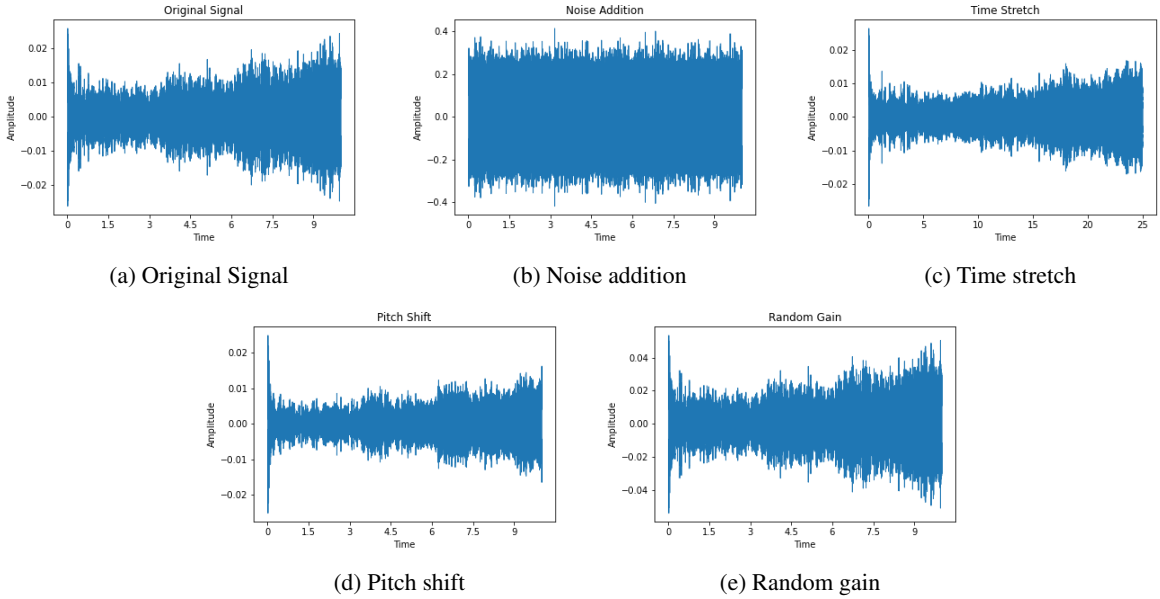


Fig. 2: Extracted MFCC features for a sample audio file

also helps in dimensionality reduction. The new reduced set of features summarises the majority of the information in the original set of features. The feature set chosen for our task is the Mel-Frequency Cepstral Coefficients (MFCCs). The motivation behind using MFCCs is that MFCCs model the vocal tract (filter) when vocal cords vibrate (impulse) to produce sound. Similarly, they can model the acoustic environment (filter) where impulse of sound (impulse) is produced. They are widely used for the purpose of speech recognition as well as for reliable audio classification. A matrix of 431 x 13 MFCCs is obtained and fed to the network. Figure 2 shows the plot of MFCC features extracted from a sample audio file.



(a) Original Signal (b) Noise addition (c) Time stretch

(d) Pitch shift (e) Random gain

Fig. 3: Data Augmentation on sample audio signal

1.4 Training

In semi-supervised learning the model training is very different from model training in normal supervised methods. We have used the Π-model in our work for semi-supervised learning which a generalization of the τ-model of the ladder network used in [16].

The detailed block diagram for calculating training loss is shown in Figure 4. It can be observed from the figure that there are 2 major loss components: Supervised loss and Unsupervised loss. The labelled data (40%) is used to compute the supervised loss whereas entire training data is used to compute the unsupervised loss. The Π-model promotes unvarying network predictions between two realisations of the same input sample (augmented and original) under two different dropout scenarios. Dropout regularisation and input augmentation are critical components of our technique. Since they introduce variations in the input samples, it provides a reliable way to train the unlabelled data.

Algorithm 1 describes the training process using the Π-model. The network is tested twice for each training input X_i during a training epoch, generating two prediction vectors Z_i and $\tilde{Z}_i$. As previously stated, the loss function consists of two parts. The first part is the normal cross-entropy loss, which is only estimated for labelled samples. The second part, which is estimated for both labelled and unlabelled samples, penalises different predictions for the same training input X_i by computing the mean square difference between Z_i and $\tilde{Z}_i$. This effectively tries to reduce the distance between two evaluations for the same input by comparing the prediction vectors Z_i and $\tilde{Z}_i$, which is a much stronger criterion than merely ensuring that the final predictions and true labels remain the same, which generally happens in the traditional supervised training. A time-dependent weighting function $w(t)$ is used to integrate the supervised and unsupervised loss components.

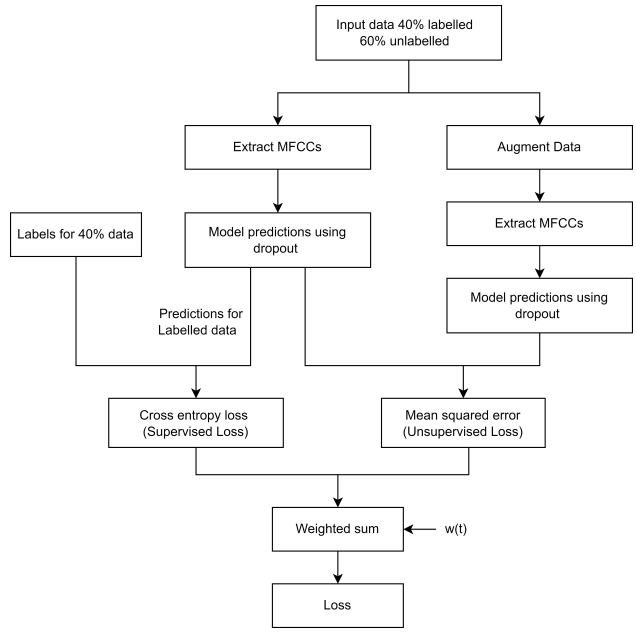


Fig. 4: Detailed block diagram for estimation of Loss

Because of the random nature of dropout regularisation, there are variations in the network predictions while training. As a result, two predictions of the same input X_i with the same network weights can provide varied results. Furthermore, augmentations add to these variations. As a result of these influences, there is a divergence between the output vectors Z_i and $\tilde{Z}_i$. Given that the initial input X_i was the same, this variance can be considered as a classification error, and so reducing it is the goal of the training process.

The unsupervised loss weighting function $w(t)$ gradually increases along a Gaussian curve starting from 0, for the first few training epochs. This ensures that at first the supervised loss component, or simply the labelled data,

Algorithm 1 Algorithm for training

Require:
Require: X_i = training samples
Require: N = total number of epochs
Require: L = training samples with labels
Require: B = training batch size
Require: C = total number of classes
Require: Y_i = labels for labelled training stimuli $i \in L$
Require: $m(x)$ = feature extraction function
Require: $a(x)$ = stochastic input augmentation function
Require: $f_\theta(x)$ = neural network with trainable parameters θ
Require: $w(t)$ = unsupervised weight ramp-up function
for j in $[1, N]$ **do**
  for i in $[1, B]$ **do**
    $Z_i \leftarrow f_\theta(m(X_i))$ // Network outputs for original inputs
    $\tilde{Z}_i \leftarrow f_\theta(m(a(X_i)))$ // Network outputs for augmented inputs and different dropout
    $Loss \leftarrow -\frac{1}{|B|}\sum_{i\in L}^{B} Y_i log(Z_i)$ // Supervised loss component
      $+w(t)\frac{1}{|B|\times C}\sum_{i}^{B}||\tilde{Z}_i - Z_i||^2$ // Unsupervised loss component
    Update θ using ADAM optimizer // Update network parameters
  end for
end for

dominates the overall loss and learning gradients. After a certain number of epochs, the weight remains constant. Reducing the rise of the unsupervised loss component is crucial; or else, the network can quickly get stuck in a degenerate solution and start diverging. The results show that, when combined with a strong convolutional network architecture, the model achieves a very high classification accuracy.

The CNN Architecture used comprises of 11 convolutional layers accompanied by global average pooling and softmax function at the output layer. 30% dropout is used at the initial layers and 50% dropout at the final layer. Adam optimizer is used to calculate gradients.

2 RESULTS AND DISCUSSION

2.1 Percent labelled data and number of epochs

The effect of percentage of labelled data on model training and accuracy is observed. Even 40% labelled data gave results at par with 100% labelled data (Supervised learning). This signifies the contribution of unlabelled data and unsupervised loss term in the model training process. This percentage can be reduced even further when the training data size is large. Hence, we can say that SSL can be used reliably for many applications.

Also, the effect of number of epochs was examined on the learning curve. Figure 5,6,7 discuss the results for different number of epochs. Table 1,2,3 mentions the best accuracies obtained during training for different number of epochs. There is no significant difference in the accuracies obtained but a training period of 80 epochs serves to be sufficient for good results. The system performance degrades a bit for 90 or more number of epochs.

Table 1: Accuracies for different percent of labelled data with a model trained for 70 epochs

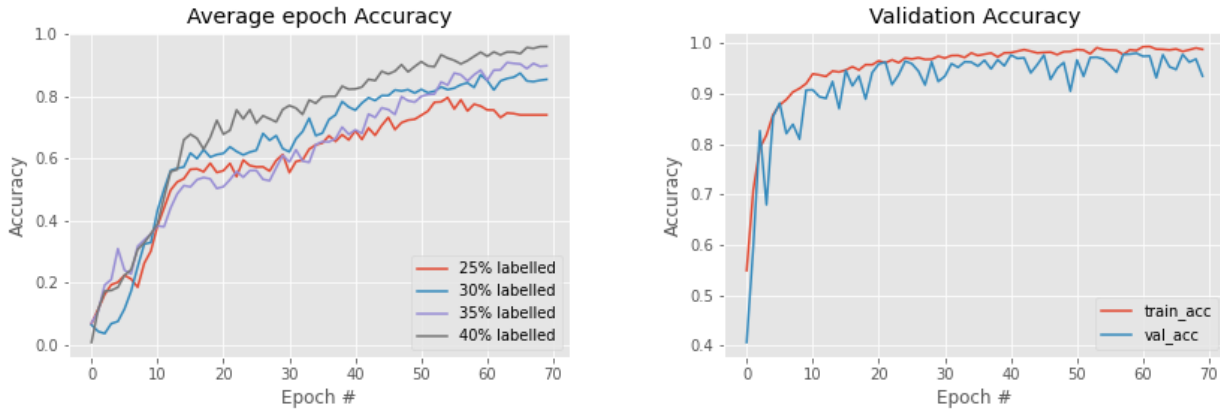
Semi-supervised Learning		Supervised Learning
Percent labelled	Accuracy	Accuracy
25	79.6%	97.9%
30	87.5%	
35	90.0%	
40	96.0%	

Table 2: Accuracies for different percent of labelled data with a model trained for 80 epochs

Semi-supervised Learning		Supervised Learning
Percent labelled	Accuracy	Accuracy
25	86.6%	98%
30	87.4%	
35	93.6%	
40	96.3%	

2.2 Pseudo Labelling

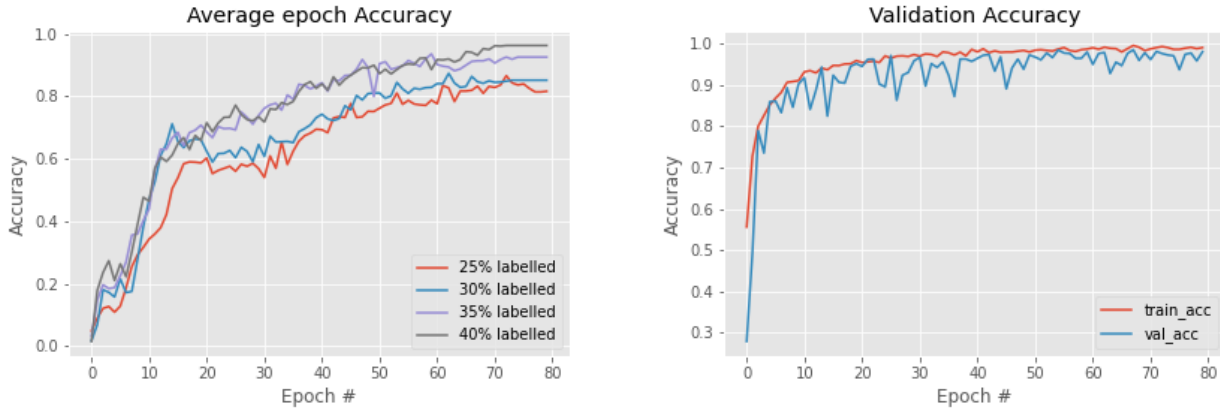
In a separate experiment, we investigated whether our approach is resistant to inaccurate labels by randomly allocating a label to a fraction of the training data before the



(a) SSL Accuracy plots for different percent of labelled data

(b) Supervised Learning Accuracy plots

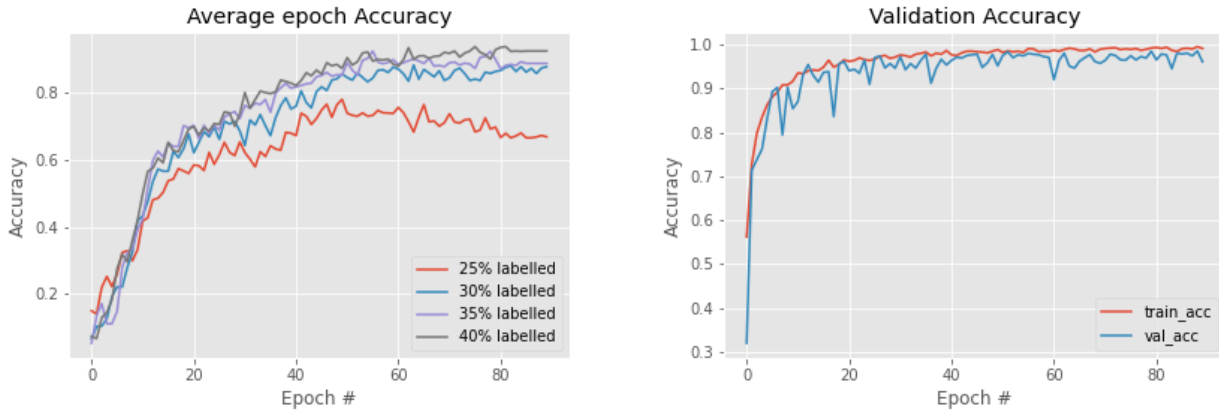
Fig. 5: Accuracy plots for 70 epochs



(a) SSL Accuracy plots for different percent of labelled data

(b) Supervised Learning Accuracy plots

Fig. 6: Accuracy plots for 80 epochs



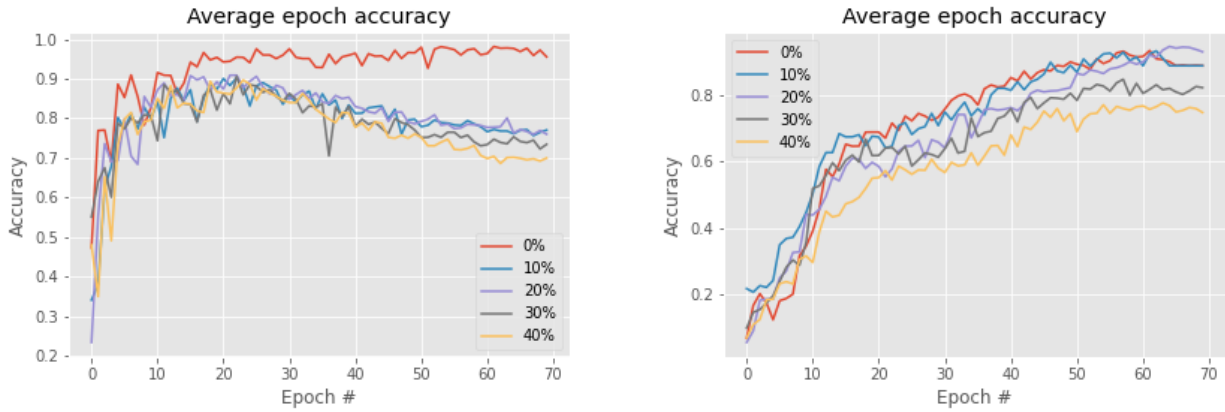
(a) SSL Accuracy plots for different percent of labelled data

(b) Supervised Learning Accuracy plots

Fig. 7: Accuracy plots for 90 epochs

training began. The classification accuracy plots for normal supervised training and Π-model are shown in Figure 8.

It is observed that due to label randomization, the accuracy for Supervised learning starts degrading after about 25 epochs whereas accuracy in case of SSL is not greatly affected due to incorrect labels. We believe that this is due to the fact that the unsupervised loss term encourages the decision boundary mapping to be generalised in the neigh-



(a) Supervised Learning

(b) Semi-supervised Learning

Fig. 8: Effect label randomization with different percent of training data on accuracy in Supervised and Semi-supervised Learning

Table 3: Accuracies for different percent of labelled data with a model trained for 90 epochs

Semi-supervised Learning		Supervised Learning
Percent labelled	Accuracy	Accuracy
25	61.7%	98.6%
30	88.4%	
35	92.5%	
40	93.9%	

bourhood of all input samples, whereas the supervised loss term forces the decision boundary to have a particular value in the neighbourhood of the labelled input samples, in simple words, it tries to overfit the data. SSL allows decision boundary to be generalised and is not fully reliant on labels in the learning process. Apart from this, it is worth noting that the learning curve for SSL increases gradually whereas that of Supervised is very steep and touches 90% accuracy within 20 epochs itself. We can say that, SSL approach is much more stable and robust as compared to the Supervised learning.

3 CONCLUSION

In the field of audio authentication immense amount of work has been done in the area of Machine Learning. However, a large part of it revolves around Supervised learning algorithms and features used. The main objective of this work was to explore Semi-Supervised Learning for audio authentication purpose. The results obtained were much the same as Supervised Learning using the same model. In fact, SSL approach tries to capture the data distribution even better. It is resistant to wrong labels. Hence, more and more data can be used to improvise the performance of the model without the laborious task of annotating data. SSL has been previously tested for image dataset [12] but not much research has been done in the audio domain. Semi-supervised learning using robust algorithms like the Π-model used in this work, proves to be efficient for audio as well.

4 FUTURE WORK

For data hungry models in Machine Learning and Deep Learning, SSL approach provides the advantage of using immense data for training without the need for manually labelling the entire data. The Π-model used can give even better results by increasing the data size. This is because Semi-Supervised Learning has a large dependency on unlabelled data to capture the most generalized boundary for best accuracy in classification. In this work about 8190 training samples were used. If this data size is increased further, the percentage of labelled data can be decreased even more.

The Π-model makes predictions based on the current epoch only. Accuracy can improve further by taking an ensemble of previous predictions for the unsupervised loss component. Besides, the model has to compute the predictions twice (once for non-augmented inputs and then for augmented inputs) in one network evaluation. This increases the computation time. There is a scope to explore more robust algorithms which require a lesser computation time in SSL.

5 REFERENCES

[1] H. Malik, H. Farid, “Audio forensics from acoustic reverberation,” presented at the *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1710–1713 (2010), [Online]. Available: 10.1109/ICASSP.2010.5495479.

[2] R. K. Patole, P. Rege, P. Suryawanshi, “Acoustic environment identification using blind de-

reverberation," presented at the *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pp. 495–500 (2016), [Online]. Available: 10.1109/CAST.2016.7915019.

[3] H. Zhao, H. Malik, "Audio forensics using acoustic environment traces," presented at the *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 373–376 (2012), [Online]. Available: 10.1109/SSP.2012.6319707.

[4] H. Zhao, H. Malik, "Audio Recording Location Identification Using Acoustic Environment Signature," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1746–1759 (2013), [Online]. Available: 10.1109/TIFS.2013.2278843.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120 (1979), [Online]. Available: 10.1109/TASSP.1979.1163209.

[6] K. Yao, K. K. Paliwal, S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Communication*, vol. 42, no. 1, pp. 5–23 (2004), [Online]. Available: https://doi.org/10.1016/j.specom.2003.09.002, adaptation Methods for Speech Recognition.

[7] R. Patil, R. K. Patole, P. P. Rege, "Audio Environment Identification," presented at the *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5 (2019), [Online]. Available: 10.1109/ICCCNT45670.2019.8944427.

[8] D. Bonet-Solà, R. Alsina-Pagès, "A Comparative Survey of Feature Extraction and Machine Learning Methods in Diverse Acoustic Environments," *Sensors*, vol. 21, p. 1274 (2021 02), [Online]. Available: 10.3390/s21041274.

[9] R. Patole, P. Rege, "A Comparative Analysis of Classifiers and Feature Sets for Acoustic Environment Classification," *Journal of the Audio Engineering Society*, vol. 67, pp. 939–952 (2019 12), [Online]. Available: 10.17743/jaes.2019.0042.

[10] M. Mulimani, S. Koolagudi, "ACOUSTIC SCENE CLASSIFICATION USING MFCC AND MP FEATURES," (2016 09).

[11] H. Malik, H. Mahmood, "Acoustic environment identification using unsupervised learning," *Security Informatics*, vol. 3, no. 1, pp. 1–17 (2014), [Online]. Available: 10.1186/s13388-014-0011-7.

[12] S. Laine, T. Aila, "Temporal Ensembling for Semi-Supervised Learning," (2016), [Online]. Available: 10.48550/ARXIV.1610.02242.

[13] A. Ghasemi, H. Rabiee, M. Fadaee, M. Manzuri, M. H. Rohban, "Active Learning from Positive and Unlabeled Data," (2016 02).

[14] A. Mesaros, T. Heittola, T. Virtanen, "TUT database for acoustic scene classification and sound event detection," presented at the *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132 (2016), [Online]. Available: 10.1109/EUSIPCO.2016.7760424.

[15] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, "DCASE 2017 challenge setup: tasks, datasets and baseline system," presented at the *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 85–92 (2017), detection and Classification of Acoustic Scenes and Events Workshop ; Conference date: 01-01-2000.

[16] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, T. Raiko, "Semi-Supervised Learning with Ladder Networks," (2015), [Online]. Available: 10.48550/ARXIV.1507.02672.

THE AUTHORS

Mrudula Jadhav

Advait Dixit

Rashmika Patole

Mrudula Jadhav is currently pursuing a B.Tech. degree in Electronics and Telecommunication Engineering from the College of Engineering, Pune, India, since 2018. Her research interests include audio processing, Image processing, and machine learning.

●

Advait Dixit is currently pursuing a B.Tech. degree in Electronics and Telecommunication Engineering from the College of Engineering, Pune, India, since 2018. His research interests include Image processing, audio processing, deep learning, and machine learning.

●

Rashmika Patole received a B.Tech. degree in Electronics and Telecommunication Engineering from the College of Engineering, Pune, India, in 2010, and M.Tech. degree in Signal Processing from the same institute. She received her Ph.D. from the Savitribai Phule Pune University, Pune, India in 2021. She has been working with the Department of Electronics and Telecommunication, College of Engineering, Pune, India, as an Assistant Professor since 2013. Her research interests include speech processing, audio authentication, audio forensics, and machine learning.