Environmental Scene Classification: Comparing Pseudolabeled Data and Traditionally Augmented Data

Ty Eaden
Lyle School of Engineering
Southern Methodist University
Dallas, Texas, United States
teaden@smu.edu

***Abstract*—Although the ESC-50 data set is widely used for audio classification endeavors, the related but weakly labeled ESC-US data set of 250,000 records has remained largely unexplored. Given that ESC-50 is a sparse data set consisting of 2000 records across 50 classes, this study investigates a method related to semi-supervised learning for utilizing ESC-US to increase ESC-50 classification performance. This method involves using a pre-trained convolutional neural network (CNN) fine-tuned on ESC-50 to create new training data by providing pseudolabels for the ESC-US records. New CNNs are then trained on subsets of ESC-50 with the addition of either traditionally augmented versions of the ESC-50 data or pseudolabeled ESC-US data. Traditional data augmentations for audio classification include pitch shifting and adding noise. The goal is to compare the impact of new pseudolabeled data and augmented versions of the original data on classification performance. The study finds that whether traditional data augmentation methods remain advantageous over new training data is highly dependent on the distribution of the new data.**

Keywords—environmental scene classification, convolutional neural networks, semi-supervised learning, pseudolabels, data augmentation

I. MOTIVATION

This investigation is important for numerous reasons. First, audio classification endeavors are often constrained by sparse data sets. ESC-50 only has 2000 records for 50 classes [1], but the problem extends beyond environmental sound classification. In terms of music genre classification, the GTZAN data set only consists of 1000 30-second audio files split across 10 different classes [2]. For speech emotion recognition, the Acted Emotional Speech Dynamic Database (AESDD) consists of 600 recorded phrases for 5 emotion classes [3]. Limited size means that data augmentation practices, such as time stretching audio clips or rotating mel spectrogram images, is necessary to achieve desirable evaluation performance. However, finding the right balance of augmentation practices for training data amounts to additional hyperparameters that must be tuned on top of model configuration criteria, such as the number of layers in the network. For example, data augmentation methods that expand a data set with only slight variations could cause a model to overfit to particular features. Although some audio data sets are much larger, as AudioSet contains 2,084,320 10-second clips across 632 labels, this unwieldy size also makes model training an arduous and lengthy process [4]. Without exceptional computing power, individuals are largely limited to using models pre-trained on AudioSet rather than training new models on AudioSet. In short, many audio classification tasks require additional means of procuring more training data effectively and efficiently.

A. Research Questions

This study seeks answers to a number of relevant questions:

- For a particular audio classification task, can uncompressed waveform data be effectively enhanced by new data of compressed formats, such as .ogg?
- For a particular audio classification task, can utilizing semi-supervised learning techniques result in confident pseudolabels for relevant but unlabeled data?
- If achieving confident pseudolabels is possible, how does training on confidently pseudolabeled data from another source compare to training on augmented versions of data from an initial source?

B. Hypothesis

The author hypothesizes that, since short environmetnal sounds may be simpler than the intracies of speech data, for example, fine tuning on .ogg files instead of uncompressed wave files will not necessarily result in performance degradation. In fact, after being fine tuned on .ogg data instead of waveform data, the model should be able to confidently pseudolabel .ogg files from outside of the training data if there is sufficient reason to believe that the outside data follows a similar distribution. Lastly, given that achieving the right balance of data augmentation methods can be a complicated process, there should be evidence that confidently pseudolabeled data from another source is linked to better classification performance than augmentations of training data from an original source.

II. RELATED WORK

This analysis expands upon the work of many other practitioenrs. Since this study utilizes the ESC-50 and ESC-US data sets, it builds on the work of scientist Karol J. Piczak, who collected the data for Harvard University in 2015 [5].

Since the goal is to maximize classification performance on the ESC-50 data set, this study references models that have already succeeded in this task. These architectures include the transformer-based InternVideo2 [6], OmniVec [7], and Beats [8], which respectively achieve classification accuracy rates of 98.6%, 98.4%, and 98.1% after fine tuning on ESC-50. As further explained in the methodology section, this experiment directly utilizes mn40_as, which is a variant of Google's MobileNetV3 CNN that achieved a classification accuracy rate of 97.45% after fine tuning on ESC-50 [9].

This study also references successful applications of data augmentation, particularly for environmental sound classification. In *Spectral images based environmental sound classification using CNN with meaningful data augmentation*, Zohaib Mushtaq, Shun-Feng Su, and Quoc-Viet Tran of the National Taiwan University of Science and Technology (NTUST) proposed their Novel Augmentation Approach (NAA). Traditional data augmentation techniques involve flipping and rotating spectrogram images after conversion from audio waveform. The NTUST research team showed that techniques like pitch shifting, silence trimming, and time stretching applied to the original audio before spectrogram conversion can lead to even better performance. Despite the small size of the data sets, the team was able to achieve 99.04% and 97.57% accuracy rates on ESC-10 and ESC-50, respectively, using the NAA approach [10].

Lastly, the author was inspired to investigate pseudolabeling unlabeled data as a means of gaining additional training records after reading the semi-supervised learning work of Siddharth Gururani and Alexander Lerch. These practitioners represent the Georgia Institute of Technology [11]. Gururani and Lerch utilized multiple teacher-to-student model learning procedures, including the Mean Teacher semi-supervised learning algorithm, to provide confident labels for the OpenMic musical instrument data set and the SONYC Urban Sound Tagging data set. Without the application of labeling processes, these data sets respectively miss 90% and 6% of their labels.

III. Methods

A. Dataset and Evaluation Metrics

The ESC-50, ESC-10, and ESC-US data consists of "5-second-long clips, 44.1 kHz, single channel, Ogg Vorbis compressed @ 192 kbit/s" from the Freesound.org project [5]. TABLE I shows the distribution of categories for both ESC-50 as well as the ESC-10 subset [12]. The ESC-50 data set has 2000 records while the ESC-10 data set has 400 records. There are 40 records per category, so ESC-50 is a balanced data set and accuracy can therefore be used as an evaluation metric. The ESC-US data set consists of 250,000 records, which are only weakly labeled by user-specified tags from Freesound.

TABLE I. ESC-50 AND ESC-10 (BOLDED)

ESC Categories				
Animals	***Natural***	***Human***	***Interior***	***Exterior***
Dog	**Rain**	**Crying Baby**	Door Knock	**Helicopter**
Rooster	**Sea Waves**	**Sneezing**	Mouse Click	**Chainsaw**
Pig	**Crackling Fire**	Clapping	Keyboard	Siren
Cow	Crickets	Breathing	Door, Creaks	Car Horn
Frog	Chirping Birds	Coughing	Can Opening	Engine
Cat	Water Drops	Footsteps	Washing Machine	Train
Hen	Wind	Laughing	Vacuum Cleaner	Church Bells
Insects	Pouring Water	Brushing Teeth	Clock Alarm	Airplane
Sheep	Toilet Flush	Snoring	**Clock Tick**	Fireworks
Crow	Thunderstorm	Drinking	Glass Breaking	Hand Saw

It is worth mentioning that an updated version of the ESC-50 data set exists that includes uncompressed audio waveform files rather than .ogg files [13]. However, given the fact that ESC-US is only available in .ogg form, the .ogg form of ESC-50 was also utilized for consistency. In advance of any model training, the ESC-50 and ESC-US .ogg files were decoded into .wav files.

B. Method 1: Pseudolabeling ESC-US with mn40_as CNN

The mn40_as CNN is a variant of Google's MobileNetV3 CNN that includes 68.43 million parameters [14]. The model is specifically a version of the proposed MobileNetV3-Large architecture. Compared to a standard feedforward CNN, MobileNetV3-Large includes the following characteristics [15]:

- Depthwise Separable Convolutional Layers: Splits convolution into filtering (i.e., depthwise) as well as combining (i.e., pointwise) operations and is very computationally efficient compared to traditional convolutional layers
- Squeeze-and-Excitation Blocks:
  - Squeeze: Pools activations over height and width dimensions
  - Excitation: Passes pooled activations through bottleneck layer that produces channel weights, which scale the pre-squeezed activations across the channel dimension
- Hard-Swish Activation Function: Provides nonlinear activation at a fraction of the computational cost of ReLU

The mn40_as CNN takes as input log mel spectrograms, and it was pre-trained on ImageNet [9]. The model was further trained on AudioSet via knowledge distillation from an ensemble of teacher Patchout faSt Spectrogram Transformer (PaSST) models. PaSST modifies the Vision Transformer (ViT) architecture to work on audio tasks [16]. This PaSST model takes spectrograms as input, which it divides into patches and sends to a transformer encoder.

The mn40_as architecture achieves 97.45% accuracy when fine-tuned on ESC-50. However, this accuracy metric is for the updated version of ESC-50, which includes uncompressed waveform files. Since ESC-US is only available in .ogg format, it makes sense to fine tune mn40_as on the decoded ESC-50 .ogg files. If the model retains similarly high accuracy on the decoded ESC-50 .ogg files, it can then be used to infer effective pseudolabels for the ESC-US data. The mn40_as softmax probabilities associated with the model predictions are used as confidence thresholds for the pseudolabels. This study evaluates the number of records associated with each pseudolabel (i.e., ESC-50 category) at 50%, 70%, and 90% confidence thresholds.

C. Method 2: Training Classification Models

A total of six classification models are trained from scratch for this study based on the aforementioned confidence thresholds.

1) Pseudolabeld Data: For the 50%, 70%, and 90% confidence thresholds, the first three models are trained on original ESC-50 data plus pseudolabeled ESC-US data. This

means that, if 37 pseudolabels were associated with records above the 50% confience threshold, the first model would be trained on a subset of the original ESC-50 data for those 37 classes in addition to the ESC-US pseudolabeled data for those 37 classes. In the original ESC-50 data, there are 40 records per class, and there were 200 pseudolabeled records chosen per class. In other words, only classes associated with at least 200 pseudolabeled records for a given confidence threshold were considered. In cases where more than 200 pseudolabeled records exist at the given confidence threshold for a particular class, the 200 records with the highest softmax confidence thresholds are selected. The target of 200 records per pseudolabel aligns with the number of per-class records generated by the subsequently discussed data augmentation strategy. Overall, the first model would be trained on 240 ESC-50 plus ESC-US pseudolabeled data records for the 37 classes. Since all classes are associated with 240 records, the data used for training is balanced.

2) Augmented Data: The second three models are respectively trained on the same subsets of ESC-50 original data. Again, if the 50% confidence threshold had 37 pseudolabels associated with record counts over 200, the first augmented model would be trained on a subset of the original ESC-50 data for those 37 classes. However, rather than being additionally trained on ESC-US data, the second three models are trained on augmented versions of the original ESC-50 data for the aforementioned classes. There are five augmentations, each of which results in 40 additional records for a given classes. Five augmentations multiplied by 40 additional records plus the original 40 records per class means that the second three models, like the first three models, are trained on 240 records per class. Using the same per-class record count for the pseudolabel models and the augmented models allows for clean comparisons. The included data augmentations are as follows [10]:

a) Positive pitch shift: Apply shift of +2

b) Negative pitch shift: Apply shift of -2

c) Slow Time Stretch: Apply slowing factor of 0.7

d) Fast Time Stretch: Apply quickening factor of 1.2

e) Trim Silence: Trim audio under 40 dB

3) Model Architecture: There are likely more classes associated with over 200 pseudolabeled records at a 50% confidence threshold than at a 90% confidence threshold. Therefore, the models assocaited with different confidence thresholds are trained on different amounts of records (although the per-class record counts remain balanced at 240). While one strategy may involve coming up with different model architectures for the different confidence thresholds to accomodate differences in record numbers, this study opts to use one model architecture for all thresholds. Holding the model architecture constant allows one to see how different amounts of pseudolabeled and augmented data impact performance in terms of overfitting and accuracy. The architecture arbitrarily chosen in TABLE II is a small VGG-like CNN model. Only four convolutional layers are used to reflect the simplicity of the 5-second environmental sounds. Batch normalization and dropout are applied in an attempt to mitigate or at least minimize instances of overfitting. The model accepts input log mel spectrograms of shape (128, 128, 1). The Adam optimizer is used with a learning rate of 1e-5. The chosen loss function is categorical cross-entropy. The model was trained on batches of 32 records for 10 epochs.

TABLE II. CNN ARCHITECTURE

Small VGG-Like CNN (4 Convolutional Layers)	
Layer Type	***Configuration***
Input Layer	input_shape=(128, 128, 1)
Conv2D	32 filters, (3x3) kernel, relu, same padding
BatchNormalization	-
Conv2D	32 filters, (3x3) kernel, relu, same padding
BatchNormalization	-
MaxPooling2D	(2x2) pool size
Conv2D	64 filters, (3x3) kernel, relu, same padding
BatchNormalization	-
Conv2D	64 filters, (3x3) kernel, relu, same padding
BatchNormalization	-
MaxPooling2D	(2x2) pool size
Flatten	-
Dense	128 units, relu activation
BatchNormalization	-
Dropout	0.5 rate
Dense	50 units, softmax activation

D. Method 3: Levene Test For Equality of Variances

Each pseudolabel and augmented model will be trained using 10-fold cross validation. The advantage of this approach is that it allows for numerous evaluation metric samples for statistical testing. An applicable statistical test is Levene's test for equality of variances. For the pseudolabel and augmented models associated with a given confidence threshold, this test can determine if respective differences in variance between training loss, validation loss, training accuracy, and validation accuracy are statistically significant. Overall, this test allows one to see whether the pseudolabeled data or the augmented data is associated with greater stability across the evaluation metrics for a given confidence threshold. For this scenario, Levene's test is preferred over Bartlett's test since Levene's test is nonparametric in that it does not assume normality among the samples. It is unlikely that samples of training loss, validation loss, training accuracy, and validation accuracy are normally distributed.

IV. RESULTS

A. Fine Tuning and Confidence Thresholds

Fig. 1 – Fig. 4 below represent the results for Method 1. Fig. 1 shows that fine tuning on decoded ESC-50 .ogg files as opposed to uncompressed wave files does not result in performance degradation. While mn40_as received a 97.45% validation accuracy score for the updated ESC-50 wave data, the model attained a 96% validation accuracy rate for the decoded .ogg files. This result means that the fine-tuned mn40_as model is capable of pseudolabeling the decoded ESC-US .ogg files with confidence.

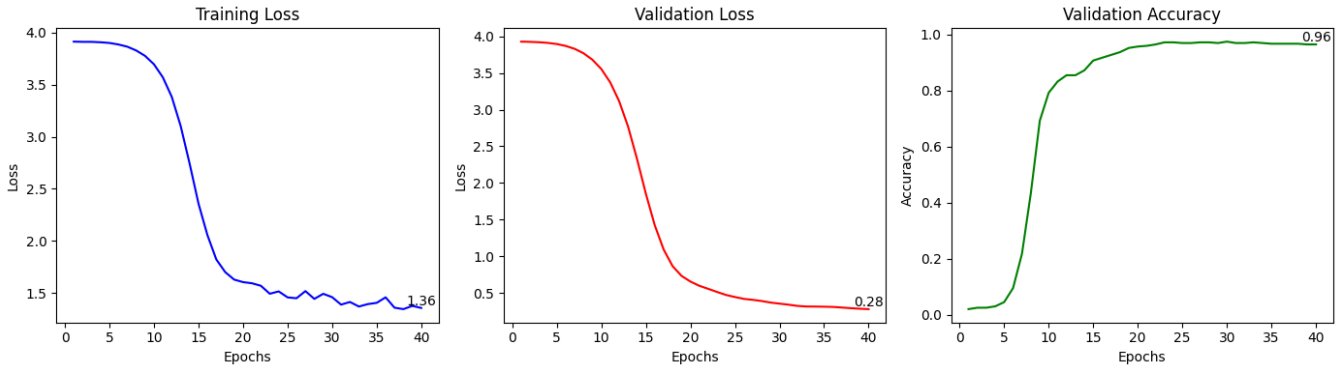

Fig. 1. mn40_as CNN Fine-Tuning: ESC-50 Decoded .ogg Files

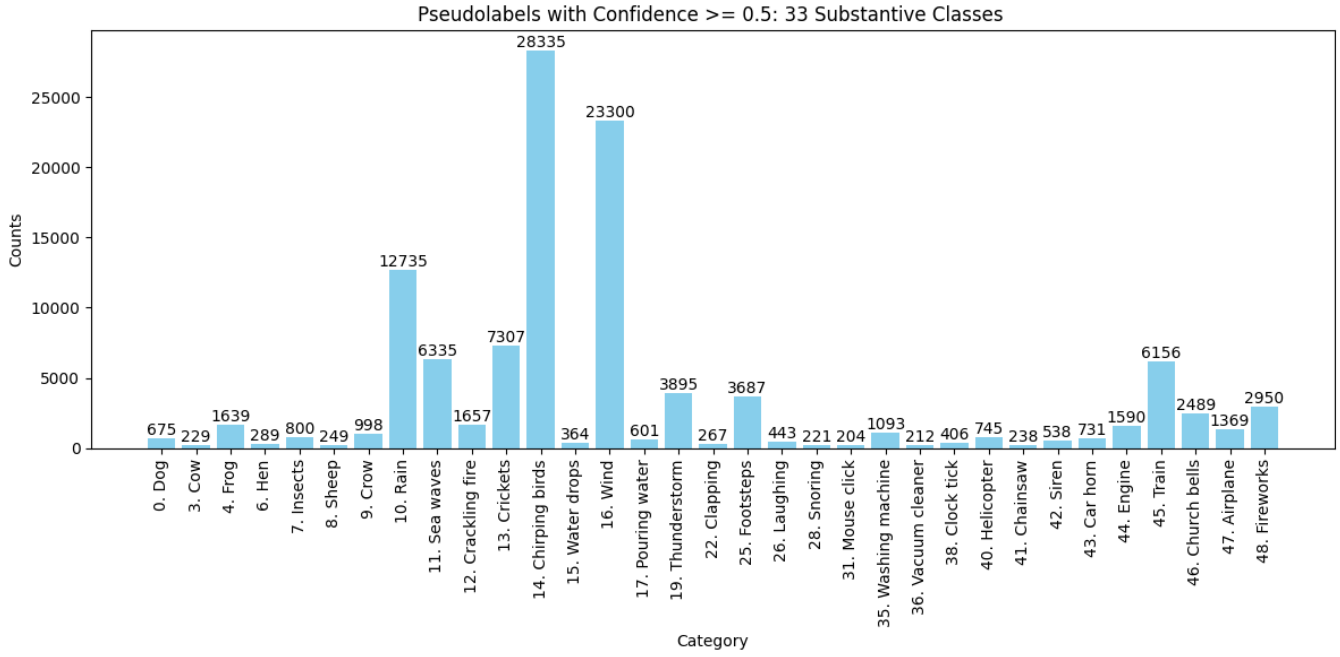

Fig. 2. ESC-50 Classes With Over 200 Pseudolabeled Records – 50% Softmax Confidence

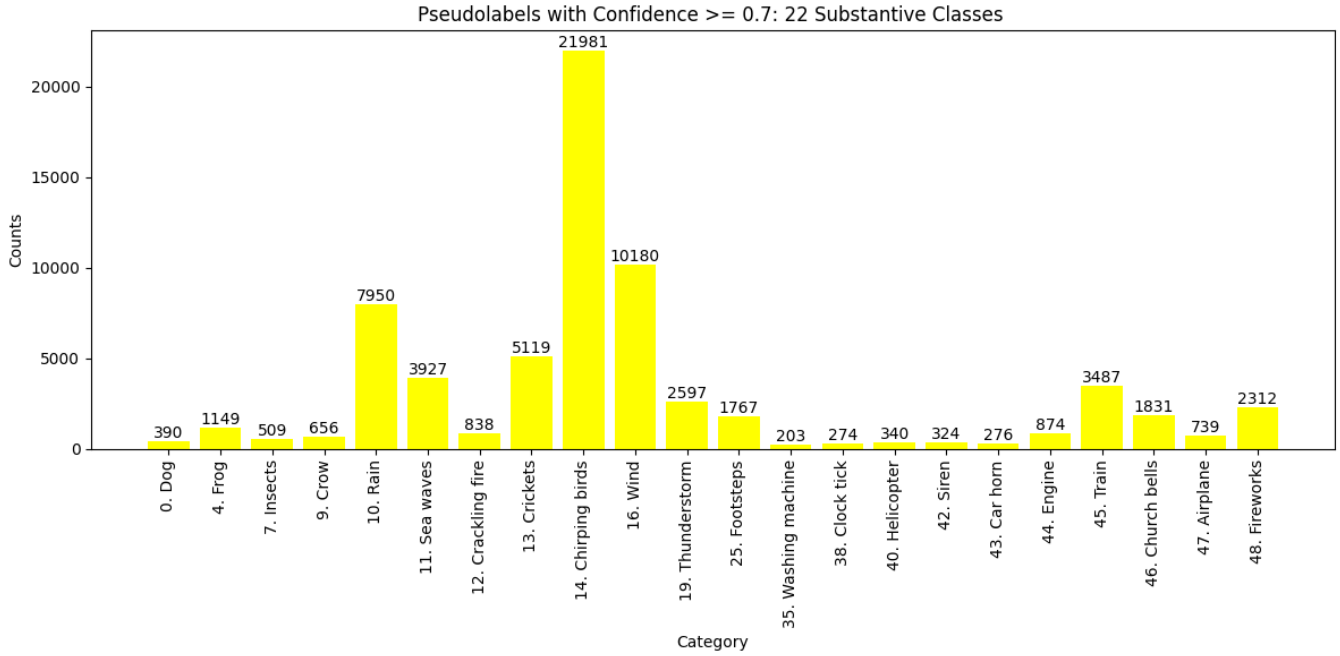

Fig. 3. ESC-50 Classes With Over 200 Pseudolabeled Records – 70% Softmax Confidence

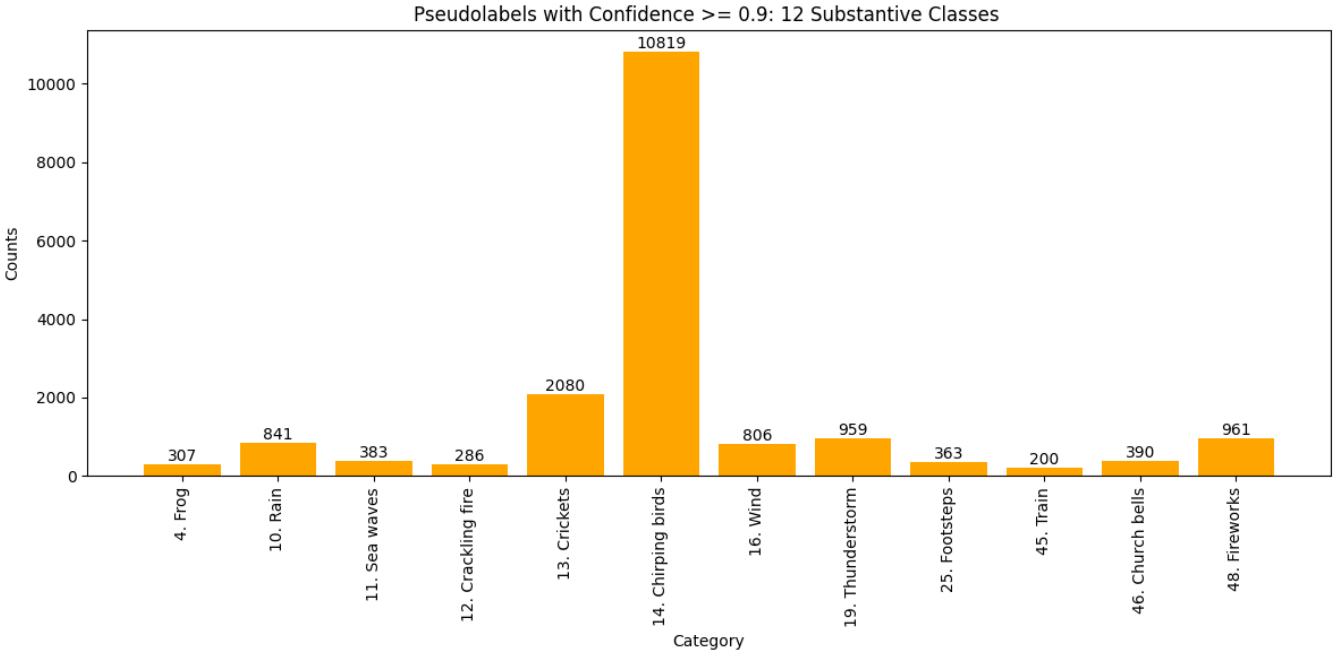

Fig. 4. ESC-50 Classes With Over 200 Pseudolabeled Records – 90% Softmax Confidence

Fig. 2 – Fig. 4 show the ESC-50 classes that are associated with over 200 pseudolabeled records for the different softmax confidence thresholds. For 50%, 70%, and 90% confidence thresholds, there are 33, 12, and 12 classes associated with greater than or equal to 200 pseudolabeled records. Compared to the models trained at 90% confidence, the models trained at 50% confidence will have more training data, but that training data is spread across a greater number of classes. Now that the number of usable classes per confidence threshold has been identified, the models trained as part of Method 2 can be defined as follows in TABLE III.

TABLE III. METHOD 2 CLASSIFICATION MODELS

Model Characteristics			
Name	***Train, Validation Description***	***Data Num***	***Threshold***
m1_33_pseudo	33 classes of original ESC-50 33 classes of ESC-US pseudolabels	7,920	50%
m2_22_pseudo	22 classes of original ESC-50 22 classes of ESC-US pseudolabels	5,280	70%
m3_12_pseudo	12 classes of original ESC-50 12 classes of ESC-US pseudolabels	2,880	90%
m4_33_aug	33 classes of original ESC-50 33 classes of augmented ESC-50	7,920	50%
m5_22_aug	22 classes of original ESC-50 22 classes of augmented ESC-50	5,280	70%
m6_12_aug	12 classes of original ESC-50 12 classes of augmented ESC-50	2,880	90%

B. Classification Model Cross Validation Results

TABLE IV – TABLE IX show training loss, validation loss, training accuracy, and validation accuracy for the six classification models across 10 cross validation folds. There are some noticeable trends. The first trend is visible for both pseudolabel and augmented models as the data set size and the number of classes decrease with increasing confidence thresholds. In general, as the confidence threshold increases, both model types see decreases in training loss and validation loss as well as increases in training accuracy and validation accuracy. This result suggests that the greater data set size for lower confidence thresholds is not enough to capture the variation among the increased number of classes. Moreover, for both pseudolabel and augmented models, the gap between training loss and validation loss as well as the gap between training accuracy and validation accuracy decrease as the confidence threshold increases.

The most notable finding is the fact that the augmented models generally outperform the pseudolabel models. If one compares any augmented model to its pseudolabel model pair within the same confidence threshold, it is clear that training and validation accuracies are higher while training and validation losses are lower for the augmented model. Additionally, while best-performing pseudolabel model m3_12_pseudo experiences overfitting based on training accuracy exceeding validation accuracy by 5%, the training and validation accuracies are nearly equivalent for augmented model m6_12_aug.

C. Levene Test For Equality of Variances Results

Fig. 5 shows the Levene variance statistics and p-values for pseudolabel-augmented model pairs within the same confidence threshold. Augmented model m4_33_aug has a statistically different training accuracy variance compared to pseudolabel model m1_33_pseudo. Model m4_33_aug has a training accuracy variance of $2.25*10^{-5}$ while m1_33_pseudo's training accuracy variance is $1.86*10^{-4}$. However, the differences in variance between respective evaluation metrics across the pseudolabel-augmented model pairs are generally not statistically significant. This indicates that training is largely stable for both the pseudolabeled data and the augmented data.

TABLE IV. M1_33_PSEUDO CLASSIFICATION METRICS

Name: m1_33_pseudo	***Validation Fold***										
Metric/Fold	1	2	3	4	5	6	7	8	9	10	Avg
Training Loss	0.76	0.83	0.77	0.74	0.76	0.73	0.70	0.76	0.75	0.64	0.74
Validation Loss	1.06	1.21	1.09	1.12	1.16	0.95	0.96	1.01	1.08	1.09	1.07
Training Accuracy	0.82	0.81	0.82	0.83	0.83	0.84	0.85	0.83	0.83	0.86	0.83
Validation Accuracy	0.72	0.70	0.71	0.72	0.72	0.77	0.76	0.76	0.72	0.71	0.73

TABLE V. M2_22_PSEUDO CLASSIFICATION METRICS

Name: m2_22_pseudo	***Validation Fold***										
Metric/Fold	1	2	3	4	5	6	7	8	9	10	Avg
Training Loss	0.45	0.48	0.44	0.47	0.50	0.45	0.46	0.47	0.44	0.46	0.46
Validation Loss	0.73	0.64	0.75	0.69	0.76	0.73	0.66	0.69	0.66	0.67	0.70
Training Accuracy	0.90	0.89	0.89	0.88	0.89	0.90	0.89	0.88	0.89	0.89	0.89
Validation Accuracy	0.79	0.81	0.80	0.79	0.78	0.80	0.83	0.81	0.81	0.79	0.80

TABLE VI. M3_12_PSEUDO CLASSIFICATION METRICS

Name: m3_12_pseudo	***Validation Fold***										
Metric/Fold	1	2	3	4	5	6	7	8	9	10	Avg
Training Loss	0.20	0.19	0.24	0.21	0.18	0.24	0.21	0.21	0.26	0.19	0.21
Validation Loss	0.32	0.31	0.41	0.36	0.30	0.38	0.35	0.28	0.37	0.32	0.34
Training Accuracy	0.95	0.95	0.94	0.95	0.96	0.94	0.95	0.96	0.93	0.96	0.95
Validation Accuracy	0.91	0.91	0.90	0.90	0.91	0.89	0.91	0.94	0.89	0.80	0.90

TABLE VII. M4_33_AUG CLASSIFICATION METRICS

Name: m4_33_aug	***Validation Fold***										
Metric/Fold	1	2	3	4	5	6	7	8	9	10	Avg
Training Loss	0.32	0.30	0.29	0.29	0.32	0.32	0.33	0.30	0.26	0.31	0.30
Validation Loss	0.37	0.28	0.31	0.34	0.35	0.34	0.36	0.37	0.27	0.32	0.33
Training Accuracy	0.96	0.96	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.96	0.96
Validation Accuracy	0.93	0.96	0.95	0.94	0.95	0.95	0.94	0.95	0.96	0.96	0.95

TABLE VIII. M5_22_AUG CLASSIFICATION METRICS

Name: m5_22_aug	***Validation Fold***										
Metric/Fold	1	2	3	4	5	6	7	8	9	10	Avg
Training Loss	0.21	0.24	0.25	0.26	0.23	0.26	0.27	0.21	0.26	0.24	0.24
Validation Loss	0.28	0.25	0.26	0.35	0.25	0.32	0.28	0.26	0.26	0.25	0.27
Training Accuracy	0.97	0.96	0.96	0.96	0.97	0.96	0.95	0.97	0.96	0.96	0.96
Validation Accuracy	0.94	0.96	0.96	0.94	0.96	0.93	0.94	0.95	0.95	0.96	0.95

TABLE IX. M6_12_AUG CLASSIFICATION METRICS

Name: m6_12_aug	***Validation Fold***										
Metric/Fold	1	2	3	4	5	6	7	8	9	10	Avg
Training Loss	0.13	0.15	0.12	0.11	0.12	0.11	0.13	0.14	0.14	0.17	0.13
Validation Loss	0.13	0.18	0.10	0.14	0.17	0.12	0.12	0.17	0.14	0.19	0.15
Training Accuracy	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.97	0.98
Validation Accuracy	0.97	0.96	0.99	0.97	0.96	0.98	0.98	0.96	0.97	0.95	0.97

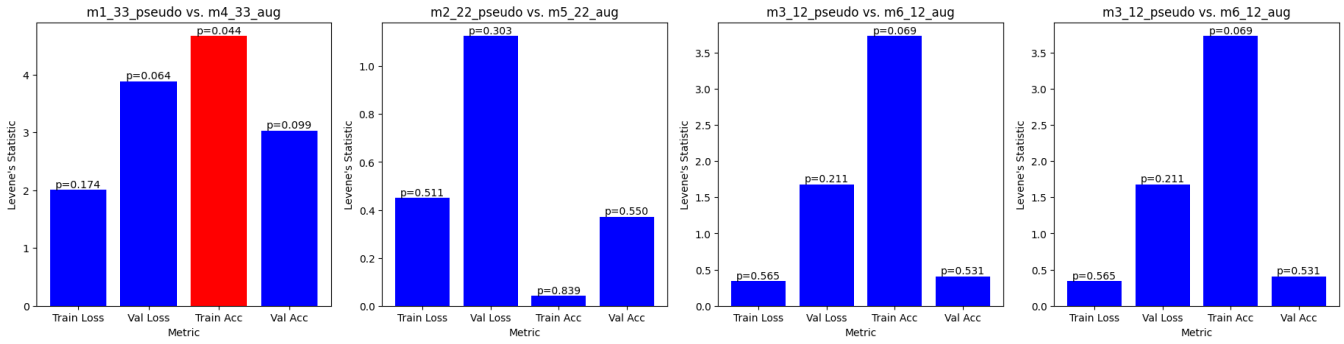

Fig. 5. Levene Variance Data For Pseudolabel And Augmented Models Within Same Confidence Threshold

V. CONCLUSION

This section revisits the research questions posed at the beginning of the paper:

- For a particular audio classification task, can uncompressed waveform data be effectively enhanced by new data of compressed formats, such as .ogg?
- For a particular audio classification task, can utilizing semi-supervised learning techniques result in confident pseudolabels for relevant but unlabeled data?
- If achieving confident pseudolabels is possible, how does training on confidently pseudolabeled data from another source compare to training on augmented versions of data from an initial source?

It was hypothesized that fine tuning on decoded ESC-50 .ogg files would result in performance comparable to fine tuning on ESC-50 uncompressed wave files. In general, the waveform files that result from decoding the .ogg files cannot regain the quality lost from compression. However, since the environmental sounds are simple and only 5 seconds in length, it was thought that compression would not significantly affect quality. This part of the hypothesis was proven true, as fine tuning mn40_as on ESC-50 decoded .ogg files results in 96% validation accuracy. This score is extremely close to the 97.45% accuracy score associated with the uncompressed wave files.

The mn40_as model also led to confident pseudolabels for the ESC-US data set. 18,395 pseudolabeled records across 12 ESC-50 classes are linked to over 90% softmax confidence. Also, 67,723 of the 250,000 ESC-US data set across 22 ESC-50 classes are linked to over 70% softmax confidence.

However, audio data augmentation, which can include pitch shifting and time stretching among other methods, led to slightly superior validation accuracies compared to the pseudolabeling strategy with decreased magnitudes of overfitting. In contrast, this study predicted that the pseudolabeling strategy would outperforming data augmentation in terms of model classification performance. After all, data augmentation strategies often require careful tuning and balancing to yield classification performance benefits. Contrary to this initial thought, the end result shows that expanding a data set by including augmentations of the original data may still yield superior results over procuring new data from a related source.

REFERENCES

[1] J. Salamon, "UrbanSound: A Dataset of Urban Environmental Sound Sources," Harvard Dataverse, V2, 2023. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YDEPUT. [Accessed: May 7, 2024].

[2] "GTZAN Genre Collection," Papers with Code. [Online]. Available: https://paperswithcode.com/dataset/gtzan. [Accessed: May 7, 2024].

[3] DagsHub, "Acted Emotional Speech Dynamic Database," 2024. [Online]. Available: https://dagshub.com/DagsHub/audio-datasets/src/main/Acted-Emotional-Speech-Dynamic-Database. [Accessed: May 7, 2024].

[4] Google, "AudioSet," Google Research. [Online]. Available: https://research.google.com/audioset/. [Accessed: May 7, 2024].

[5] J. Salamon, "UrbanSound Dataset," Harvard Dataverse, V2, 2023. DOI: 10.7910/DVN/YDEPUT.

[6] Y. Wang et al., "INTERNVIDEO2: SCALING VIDEO FOUNDATION MODELS FOR MULTIMODAL VIDEO UNDERSTANDING." Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/pdf/2403.15377v1

[7] S. Srivastava and G. Sharma, "OmniVec: Learning robust representations with cross modal sharing," 2023. Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/pdf/2311.05709v1

[8] S. Chen et al., “BEATS : Audio Pre-Training with Acoustic Tokenizers.” Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/pdf/2212.09058v1

[9] F. Schmid, K. Koutini, and G. Widmer, “EFFICIENT LARGE-SCALE AUDIO TAGGING VIA TRANSFORMER-TO-CNN KNOWLEDGE DISTILLATION.” Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/pdf/2211.04772v3

[10] Z. Mushtaq, “Spectral images based environmental sound classification using CNN with meaningful data augmentation,” www.academia.edu, Accessed: May 08, 2024. [Online]. Available: https://www.academia.edu/115328331/Spectral_images_based_environmental_sound_classification_using_CNN_with_meaningful_data_augmentation?uc-sb-sw=5486716

[11] S. Gururani and A. Lerch, “Semi-Supervised Audio Classification with Partially Labeled Data.” Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/pdf/2111.12761YouTube, "Video Title," YouTube, 2024. [Online]. Available: https://www.youtube.com/watch?v=a1G1n_kTo8s. [Accessed: May 7, 2024].

[12] K. Piczak, "ESC-50: Dataset for Environmental Sound Classification," GitHub repository, 2024. [Online]. Available: https://github.com/karolpiczak/ESC-50. [Accessed: May 7, 2024].

[13] F. Schmid, "EfficientAT," GitHub repository, 2024. [Online]. Available: https://github.com/fschmid56/EfficientAT. [Accessed: May 7, 2024].

[14] H. Howard et al., "Searching for MobileNetV3," arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/1905.02244. [Accessed: May 7, 2024].

[15] OpenAI, "ChatGPT-4," OpenAI, 2024. [Online]. Available: https://openai.com/chatgpt. [Accessed: May 7, 2024].

[16] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, “Efficient Training of Audio Transformers with Patchout.” Accessed: May 08, 2024. [Online]. Available: https://arxiv.org/pdf/2110.05069