Assessing an Open Set Classification Method to Improve Sound Anomaly Detection in Acoustic Modalities

Salena Kha[1], David Wood[2], Dinesh Verma, PhD[2]
Horace Greeley High School[1]; IBM Thomas J. Watson Research Center[2]

Abstract

In today's technological era, the ability to detect anomalies is crucial for maintaining the integrity of complex systems. This paper addresses the challenge of anomaly detection, focusing on the classification and identification of sounds to identify anomalies within sound data. Used as a general anomaly detection method, efficient Out-of-Distribution (OOD) detection is used in fields such as ecosystem intrusion detection, security anomaly detection, fraud detection, and system health monitoring. Traditional machine learning models often struggle when confronted with unseen classes while frequently occurring in dynamic environments. In this paper, we propose and assess an open-set classification approach to mitigate this issue and improve the adaptability of AI models to handle previously unseen or unknown labels within audio data. In this study, we invent a classifier that makes use of a KNN model and test a set of labeled sounds for outlier detection against groups of 5, 10, 15, and 20 models. We find that this approach to anomaly detection is successful due to the classifier detecting at least one unknown value within the set of labels. We conclude that future research should enhance the number of models trained—with a larger number of compared sound labels, the recall value is expected to increase.

Introduction

Overview of Research Problem

Industry 4.0 is a movement consisting of rapid advancements and rising manufacturing productivity in the technological field. It stands for the Fourth Industrial Revolution, where a new definition of organization and control over increasing individual customer requirements consists of the Internet of Things, Industrial Internet, Smart Manufacturing, and Cloud-based Manufacturing [1]. Significant efforts have been directed towards anomaly and outlier detection in Out-of-Distribution (OOD) data, a critical task when deploying models in real-world environments [2]. OOD training is an effective method to involve auxiliary data in training, and the quality of these samples should be similar to in-distribution (ID) to teach the model [3].

A central challenge in this domain is the classification and identification of sound data, with the goal of identifying anomalies within the data. Financial fraud, malicious activity, intrusions, and system breakdowns are examples of anomalies [4] in industrial systems. An inherent problem in artificial intelligence (AI) systems is their limited ability to recognize only the classes present in their training data; when new classes are introduced, these AI models often fail to identify both seen and unseen data. This poses a critical limitation to their utility in dynamic environments [5], becoming inefficient in a fast-changing and exponentially growing field [6]. This paper sought to address the issue of inefficient anomaly detection for auxiliary out-of-distribution data.

Anomalous Data

In domains of data analysis and machine learning, anomalous data represent a unique category of observations that distinguish themselves not solely through statistical deviations but also through their confirmed significance within the broader context of a dataset. The process of detecting anomalous data extends beyond the initial identification of outliers using mathematical tools. It involves assessing the relevance of these outliers, which may lead to the discovery of rare events or patterns that warrant further investigation.

Outlier Data

This study held the fundamental assumption that the model trained on a comprehensive spectrum of sounds and detected unknown, outlier sounds. The term "unknown" implies a lack of familiarity or conventional classification but does not necessarily imply "anomaly." It is essential to differentiate between this outlier sound data and anomalous data, as this study prioritized the recognition and categorization of only outlier sounds that deviate from the expected, typical auditory patterns.

Labeled Data Usage

The high quality of labeled data is essential as well to represent diverse scenarios, conditions, and variations present in the problem domain for models to learn [7]. Labeling is essential for supervised machine learning [8] as it is primarily used as a foundation for model training; techniques like crowdsourcing, active learning, and semi-supervised learning are used, but they may not always provide accurate and reliable labels for all industrial machine learning scenarios [9]. One of the primary challenges in addressing this study's problem lies in the limited availability of anomaly data for training [10]—a large amount of data sufficient for training will allow the model to learn more efficiently. The scarcity of appropriate data is a frequent data science obstacle in industrial settings. Our experiment's convenient reliance on labeled, non-anomaly data further stresses this challenge as an integral feature. While there is an abundance of labeled, normal data in real-world scenarios, labeled anomalous data is relatively scarce due to the rarity of anomalies occurring in real-world scenarios; this results in datasets commonly avoiding anomalies.

To overcome this larger challenge of obtaining anomalous data, this experiment leverages labeled data to declare data as outliers. This experiment uses a labeled dataset to train a model that recognizes sounds that are not present in the training dataset as an outlier.

Review of Literature

Extensive research has been conducted in the domain of Out of Distribution Detection (OOD), which shares similarities with this study's exploration of nearest-neighbor distance [11]. Existing work explores neural networks, extracting signals that traverse between network layers to identify outliers within data intervals [12]. Open-World Learning (OWL), an essential concept, focuses on a model's ability to adapt to new data [13] over time by dynamically updating itself to accommodate changes in known classes [14] and creating new known classes [15]. Zero-Shot-Learning (ZSL) aims to classify data into classes not seen during training by utilizing external knowledge or attributes associated with classes. These concepts are explored within this study in detecting anomalies within the set of sound classes.

Supervised Learning Algorithms

Alternative supervised learning algorithms, such as Gaussian Mixture Models (GMM) [16] and K-Nearest Neighbor Models (KNN) [17], have been utilized to address similar challenges in classification and detection methods. GMM considers an entire vector of features and relies on spectrograms [18] for classification [19] based on Gaussian distributions of clusters. On the other hand, KNN offers simplicity by effectively utilizing multidimensional distances, including spectrogram-based features [20], to make classifications or predictions about the grouping of an individual data point.

Problem and Objective

The focus of this study was narrowed to address the limited effectiveness of machine learning detection, specifically in the context of audio data, which includes various acoustic modalities and data types such as images and acoustics.

The primary objective was to develop an open-set classification approach optimized for the efficient identification of unknown sounds within a set of predefined classes. The challenge this study addressed arises when confronted with a previously unseen sound and the need to determine whether it belongs to one of the known classes or is distinct from all of them. When a new sound emerged, this classifier attempted to assign it to one of the predefined classes and accurately label the sound.

The underlying premise was that by exposing each classifier to data from all other classes during training, it was more inclined to provide a response that aligns with the broader category of "other sounds." This experiment aimed to assess whether this approach enhanced the classifier's ability to correctly identify unknown sounds and distinguish them from the established classes, ultimately contributing to the field of sound classification and anomaly detection.

Hypotheses

H1) By implementing this open-set classification approach, particularly in scenarios with limited anomaly-labeled data, the model will successfully recognize at least one sound as unknown, concluding an effective classification within each set.

H2) The relationship between accuracy, measured by precision and recall value, and the N number of models trained is also assessed. By increasing the number of sound labels, the recall value will increase, concluding a more accurate detection.

Methodology

Overview

This section outlines the methods employed in the experiment and key parameters that were manipulated to assess the performance of the SoundAnomalyClassifier.

Experience and Student-Mentor Roles

The student contributions included the development of the new classifier and alterations after receiving feedback for a more efficient “top-down” approach. Both student and mentor searched for and evaluated the fit of online datasets for this project. The student formalized data processing by learning the AI-Signal-Processing (AISP) [21] command-line tools while ensuring proficiency. Lastly, a procedure was

collaboratively developed to partition and curate the dataset for evaluation, and the student developed the final Bash script from this for training and testing purposes with efficiency corrections.

Unfamiliar code issues related to the larger open-source project were resolved by the mentors. Within the larger project, the mentor provided the student with AISP tools allowing the student to classify, train, test, and evaluate in proper format. The mentors provided the student with educational discussions on engineering and collaboration workings and navigations through unfamiliar software.

Experiment Overview

The primary objective of this experiment was to evaluate the model's ability to identify new, unknown sounds using the larger predetermined dataset of sounds. Due to the lack of previous studies attempting this approach, this objective was defined as successful if the model correctly classifies its held-out sounds from the training set as unknown sounds. The number of seen classes was varied by separating 5, 10, 15, and 20 labels to assess the impact of adjusting the number of N models.

Model Evaluation

For the purpose of this study, an open-source data set of environmental sounds was collected [22]. The dataset consists of WAV files sampled at 16KHz for 50 different classes. Each class corresponds to 40 raw audio samples of 5 seconds each. For sounds A, B, and C, we assess the performance of C by defining the training set as A and B and the test set as some A, some B, and the unknown C. The model is then trained on sounds A and B and finally tested against the A's, B's, and unknown C's. This process was repeated for all sounds as unknown.

Implementation

In the original case, given N number of sounds joined to N classes, a classifier was trained to assign one unknown sound to one of these N classes. However, this approach had limitations, as it would potentially misclassify a novel sound as one of the N pre-defined classes. This was not the desired outcome.

In this approach, there was a fundamental assumption that when training N classifiers, a single classifier will train label 1 as *known* and label 2-N as *unknown*. If another unknown sound label is added, the model assumes that the new sound is part of 2-N *unknown* labels.

The revised strategy addressed the shortcomings of the original approach. Instead of training a single classifier for N classes, we chose to create N individual classifiers containing K sounds. Each classifier is designed to recognize a specific sound, such as Sound 1, Sound 2, Sound K, and so on.

A. Classifier 1: This classifier is trained with one label as the unknown sound and all others as known. Another classifier is trained with the original labels. The model then may declare the sound as unknown.
B. Classifiers 2, 3, …N: If all N classifiers declare the sound as unknown, then it is unknown. If it is not declared as unknown, then the other classifier, trained on the original labels, is used to determine which of the known labels should be assigned to the unknown sound. It will categorize this sound as something other than the unknown, all other sounds 2-N.

The rationale behind this approach was that by providing each classifier with data from all other classes during training, it would be more likely to produce a conclusion closer to the "2 through N" sounds when faced with an unknown sound. This is precisely what our experiment aimed to assess. The experiment's outcome would shed light on whether our revised model was better suited to handle new, previously unseen sounds and distinguish them from the known classes.

Assessing the Relation Between Number of Models and Performance

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$F1\ Score = \frac{TP}{TP + 1/2(FP + FN)} \quad (3)$$

Based on Hypothesis 2, each model's evaluation is assessed for the relationship between the recall value in Equation 1 and the number of sounds compared. Precision value assesses the model's correct predictions over the total predictions made, and F1 evaluates the overall performance by taking the harmonic mean of Precision and Recall.

By creating a loop to train among 5, 10, 15, and 20 models, other variables are unchanged. For instance, a sound is set to unknown and then assessed among five other sounds. Here, the classifier determines if the sound is one of the five other sounds or an unknown value. Initially, each full model and sub-model was trained with an instance of the Gaussian Mixture Model (GMM). However, the dataset used in this study did not have a fine enough time resolution for each sound. To resolve inaccuracies from this, the sub-model was changed to train on the instance of a Merge KNN model for higher efficiency in terms of the location of the sound. The GMM classifier was more sensitive to time resolution, creating less accurate predictions, while this LPKNN classifier is not at all sensitive to time dependencies. When the spectrums

and spectrograms were generated for necessary sounds, it was confirmed that the LPKNN classifier compressed the feature vector of each spectrogram.

The training time of these models was an essential aspect—A notable feature is that the algorithm did not inherently have an O(n^2) training time. Rather, the perceived quadratic training time stemmed from doubling the training data for the evaluations performed. When the training data was split into subsets, the overall length of the training data time remained constant. From an algorithmic perspective, the model is trained only once for each label, so it's correctly described as O(n) in terms of the number of models. However, in practical terms, the way the training was executed resulted in an O(n^2) time complexity due to the increased number of evaluations performed with sub-models. Therefore, while the actual training time can vary depending on the approach taken, the fundamental algorithmic complexity remains O(n) in relation to the number of models. Additionally, as the number of models increased, the classification time increased linearly. The efficiency of O(n) in classification ensured that the model could swiftly assess and categorize data across varying numbers of models.

Lastly, a Bash Script program was created to obtain all proper overall precision, recall, and F1 scores as well as unknown precision, recall, and F1 scores of each evaluated model. These values were then averaged to evaluate the overall performance of the overall model, as well as to evaluate the performance of the model in detecting unknown sounds.

Results and Discussion

Overview and Example

A sample sound Airplane provides the confusion matrix shown in Figure 1 after training the model against five other sounds. This matrix shows the accuracy in predicting the sound as each label: Breathing, Brushing Teeth, Can Opening, Car Horn, and unknown.

THIS SECTION WAS INTENTIONALLY LEFT BLANK

Label Airplane Among 5 Models Confusion Matrix

Predictions	Breathing	Brushing Teeth	Can Opening	Car Horn	Unknown
Breathing	27	3	4	6	0
Brushing Teeth	0	38	1	1	0
Can Opening	0	6	34	0	0
Car Horn	0	1	2	37	0
Unknown	5	0	2	25	8

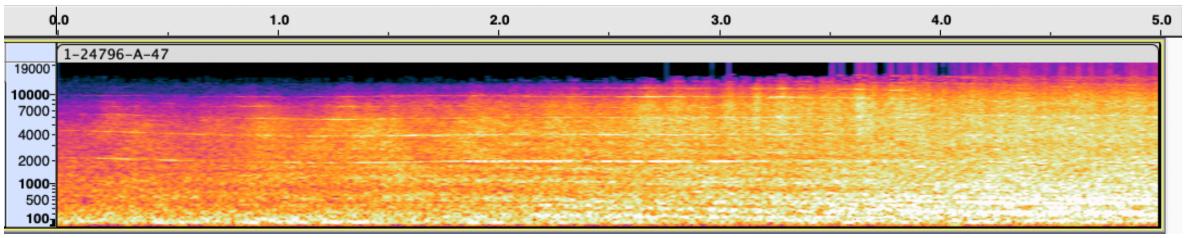
Number of Predictions

Figure 1. Label Airplane Among 5 Models Confusion Matrix

In Figure 1, the model identified the Airplane sound as unknown eight times, while the model identified it as Brushing Teeth 38 times. This example received an unknown recall score of 45.000 to represent the model's ability to recognize the sound as unknown.

Figure 2 represents the spectrogram for one sample of sound Airplane. A spectrogram is a visual representation of sound that shows how the frequency content of a sound evolves over time. This is achieved by breaking down the audio signal into small, overlapping segments and then calculating the frequency content of each segment. The resulting spectrogram provides a valuable tool for sound analysis. The model analyzed the spectrogram of the airplane sound and looked for distinctive patterns or features that matched the airplane samples it was trained on.

2a)



2b)

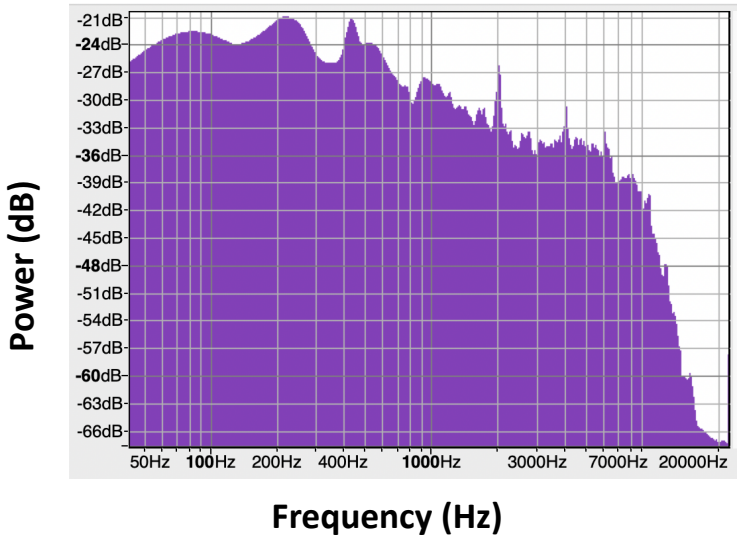


Figure 2. Sample of Airplane Spectrum (a) and Spectrogram (b)

This sample spectrum shows Power (dB) arranged by Frequency (Hz) of the single audio. In the complete dataset, there are 20 samples of an Airplane sound that trains the model to recognize its features. The model recognizes the features of each sound to capture the time-frequency characteristics of the audio.

Evaluation of Known vs. Unknown Recall, Precision, F1 Scores

Although the accuracy of overall known values had lower priority in this experiment and our focus lay on unknown sound detection, both performances were evaluated. Figures 3a, 3b, 3c, and 3d show the recall, precision, and F1, the overall average for each value across the assessed number of models, as well as the URecall, Uprecision, and UF1, which indicated the average result in detecting the unknown among 5, 10, 15, and 20 models. Focusing on the performance of unknown detection, the graphs include the Unknown Recall value averages, for 5, 10, 15, and 20, with respective values of 30%, 37%, 25.667%, and 27.625%.

THIS SECTION WAS INTENTIONALLY LEFT BLANK

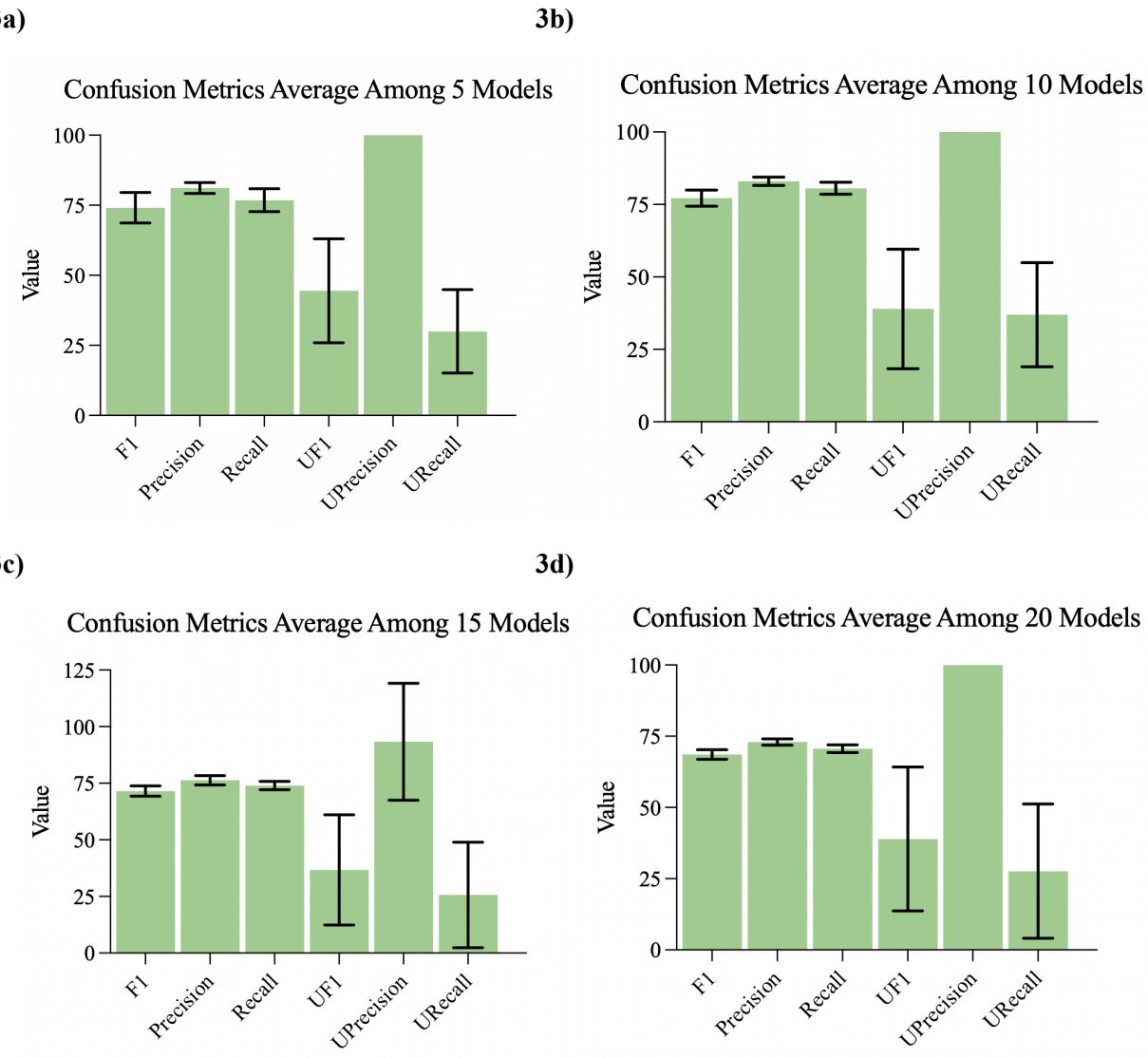


Figure 3. Overall Average Performance and Unknown Average Performance Metrics for 5 (3a), 10 (3b), 15 (3c), and 20 (3d).

The model's ability to detect any unknown values is a fundamental success for the classifier, highlighting its initial level of reliability. The precision having a frequent value of 100% is a notable feature, as there are no false positives detected; this is due to the model's implementation relying on true labels to make predictions. For example, among sounds A, B, and C, the model never detected D as an unknown. The unknown was only detected once no other sound was declared as the relevant sound. Although the model would not detect a false positive value for unknown, it may have detected a false positive for other labels, which changes the Precision value.

Re-evaluation of Group 15 Spectrograms

Figure 3 showed that Breathing is the only sound with a UPrecision value of 0. In its evaluation, the model most predicted this sound to match the Clapping sound with the highest confidence of 0.9645. Figure 4 presents both sounds to visually compare their spectrums.

4a) **4b)**

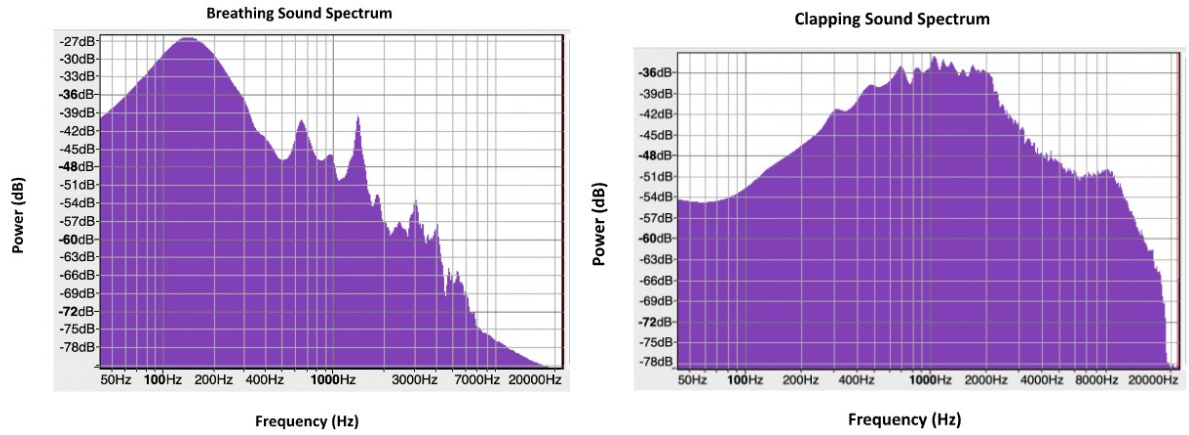


Figure 4. Breathing (a) and Clapping (b) Sound Spectrogram

The average recall value of this prediction for Breathing was 100% in three of four tests. However, visually comparing the sounds in Figure 4, their spectrums did not present major similarities, resulting in an indefinite reason for the model's prediction.

Conclusion

This study gathered a dataset that included 2,000 raw audio samples, each lasting 5 seconds, to train the model on 50 distinct sounds. Upon assessing the model's performance, it was evident that the model demonstrated the fundamental ability to detect any unknown sounds, aligning with the initial Hypothesis 1. However, the model's performance did not show a high success, particularly with lower recall averages for unknown values shown in Figure 3, including a low of 25.667% among 15 models and a high of 37% among 10 models. This is an essential aspect of the discussion as it outlines the model's current limitations in detection. The recall value was successful for detecting known values, with a high of 80.6% among 10 labels and a low of 70.631% among 20 labels. Therefore, Hypothesis 2, claiming a positive linear relationship between the number of models and recall value, was not inherently proven due to the inconsistency in recall value within all sub-groups.

However, a valuable feature was found within the precision detecting unknown values; This value remains consistently high at 100% for detection among 5, 10, and 20 labels. This characteristic is overall

advantageous in situations where precision holds greater importance than recall, most notably in the field of healthcare monitoring. In these scenarios, avoiding false positives is essential to avoiding substantial costs from unnecessary tests and treatments.

A potential improvement to enhance the model's performance is to increase the number of models trained. Increasing the number of labels to 100 or greater may yield a better recall value, allowing the model to more effectively identify unknown sounds. This study was able to improve outlier detection in acoustic modalities by inventing a classifier, which makes use of an LPKNN model, that detects unknown sounds with high precision, and is especially useful in scenarios where false positives must be avoided.

References

[1] Vaidya, S., Ambad, P. M., & Santosh Bhosle. (2018). Industry 4.0 – A Glimpse. Procedia Manufacturing, 20, 233–238. https://doi.org/10.1016/j.promfg.2018.02.034

[2] Yang, T., Huang, Y., Xie, Y., Liu, J., & Wang, S. (2022). MixOOD: Improving out-of-distribution detection with enhanced data mixup. ACM Transactions on Multimedia Computing, Communications, and Applications, 19(5), 1–18. https://doi.org/10.1145/3578935

[3] Cui, P., & Wang, J. (2022). Out-of-Distribution (OOD) Detection Based on Deep Learning: A Review. Electronics, 11(21), 3500–3500. https://doi.org/10.3390/electronics11213500

[4] Sergey Kibish. (2018, April 10). A note about finding anomalies - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/a-note-about-finding-anomalies-f9cedee38f0b

[5] Cavallaro, C., Cutello, V., Pavone, M., & Zito, F. (2023). Discovering anomalies in big data: a review focused on the application of metaheuristics and machine learning techniques. Frontiers in Big Data, 6. https://doi.org/10.3389/fdata.2023.1179625

[6] Agrawal, B., Tomasz Wiktorski, & Rong, C. (2016). Adaptive Anomaly Detection in Cloud Using Robust and Scalable Principal Component Analysis. https://doi.org/10.1109/ispdc.2016.22

[7] Laith Alzubaidi, Bai, J., Aiman Al-Sabaawi, José Santamaría, Albahri, A. S., Bashar, Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., Duan, Y., Abdullah, A., Farhan, L., Lü, Y., Gupta, A., Albu, F., Amin Abbosh, & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. Journal of Big Data, 10(1). https://doi.org/10.1186/s40537-023-00727-2

[8] Fredriksson, T., Mattos, D. I., Bosch, J., & Olsson, H. H. (2020). Data labeling: An empirical investigation into industrial challenges and Mitigation Strategies. Product-Focused Software Process Improvement, 202–216. https://doi.org/10.1007/978-3-030-64148-1_13

[9] Ivars Namatēvs, Kaspars Sudars, & Inese Poļaka. (2019). Automatic data labeling by neural networks for the counting of objects in videos. Procedia Computer Science, 149, 151–158. https://doi.org/10.1016/j.procs.2019.01.118

[10] R. Stuart Geiger, Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). “Garbage in, garbage out” revisited: What do machine learning application papers report about human-labeled training data?. Quantitative Science Studies, 2(3), 795–827. https://doi.org/10.1162/qss_a_00144

[11] A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications | ACM Computing Surveys. (2022). ACM Computing Surveys (CSUR). https://dl.acm.org/doi/10.1145/3502287

[12] Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-Distribution Detection with Deep Nearest Neighbors. PMLR, 20827–20840. https://proceedings.mlr.press/v162/sun22d.html

[13] Esmaeili, F., Cassie, E., Phan, H., Natalie, Unsworth, C. P., & Wang, A. (2023). Anomaly Detection for Sensor Signals Utilizing Deep Learning Autoencoder-Based Neural Networks. Bioengineering, 10(4), 405–405. https://doi.org/10.3390/bioengineering10040405

[14] Open-world Machine Learning: Applications, Challenges, and Opportunities | ACM Computing Surveys. (2023). ACM Computing Surveys. https://dl.acm.org/doi/10.1145/3561381

[15] Li, A., Qiu, C., Kloft, M., Smyth, P., Rudolph, M., & Mandt, S. (2023). Zero-Shot Batch-Level Anomaly Detection. ArXiv.org. https://arxiv.org/abs/2302.07849

[16] Zhou, Z. (2022). Open-environment machine learning. National Science Review, 9(8). https://doi.org/10.1093/nsr/nwac123

[17] Ramin Ranjbarzadeh, Shadi Dorosti, Saeid Jafarzadeh Ghoushchi, Caputo, A., Erfan Babaee Tirkolaee, Sadia Samar Ali, Zahra Arshadi, & Bendechache, M. (2023). Breast tumor localization and segmentation using machine learning techniques: Overview of datasets, findings, and methods. Computers in Biology and Medicine, 152, 106443–106443. https://doi.org/10.1016/j.compbiomed.2022.106443

[18] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 4(11), 218–218. https://doi.org/10.21037/atm.2016.03.37

[19] Ramakrishnan, K. (2020, July 13). Linear Algebra Vectors for Data Science - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/linear-algebra-i-vectors-for-data-science-part-i-7fe1cc5e5935

[20] Widdows, D., Kitto, K., & Cohen, T. (n.d.). Quantum Mathematics in Artificial Intelligence. https://arxiv.org/pdf/2101.04255.pdf

[21] Enterprise-Neurosystem. (2023, September 6). GitHub-Enterprise-Neurosystem/ai-signal-processing: Framework, Tools and Solutions for processing high-rate signal data such as acoustics, vibrations, etc. GitHub. https://github.com/Enterprise-Neurosystem/ai-signal-processing

[22] Moreaux, Marc. (2018). Environmental Sound Classification 50. Kaggle.com. https://www.kaggle.com/datasets/mmoreaux/environmental-sound-classification-50