

Generalization Analysis of YAMNet-DNN Architectures in Deepfake Audio Classification

Hakam Dzakwan Diash¹, Dwi Arman Prasetya², Alfian Rizaldy Pratama³, Tresna Maulana Fahrudin⁴

^{1,3}Department of Data Science, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia

²Department Master of Information Technology, University of Pembangunan Nasional Veteran Jawa Timur, Indonesia

⁴Department of Information and Communication Systems, Okayama University, Okayama 700-8530, Japan

Article Info

Article history:

Received Mar 11, 2026

Revised Apr 09, 2026

Accepted Apr 14, 2026

Keywords:

Deepfake Audio

YAMNet Embeddings

Deep Neural Networks

Domain Generalization

Fine-Tuning Strategies

ABSTRACT

The rapid advancement of speech synthesis and voice conversion technologies has increased the risk of audio deepfake attacks, necessitating robust and generalizable detection systems. This study proposes a deepfake audio detection framework that leverages pretrained YAMNet embeddings as a feature extractor, combined with statistical aggregation (Mean, Standard Deviation, and Mean + Standard Deviation) and Deep Neural Network classifiers to form compact utterance-level representations without temporal modeling. Experiments on a public dataset show that Mean and Mean + Standard Deviation representations achieve classification accuracy of up to 99.24%, outperforming Standard Deviation-based features, while deeper architectures provide only marginal performance gains. However, cross-dataset evaluation on a newly collected primary dataset reveals a significant generalization gap, with accuracy dropping to approximately 62% under direct testing. To address this issue, fine-tuning strategies with and without layer freezing are explored under limited target-domain data conditions, where the best-performing model achieves an accuracy of 0.93 on the primary dataset and 0.91 on the secondary dataset, with macro-averaged precision, recall, and F1-score reaching 0.93 and 0.91, respectively. Further evaluation on external datasets demonstrates that the model benefits from fine-tuning even without prior exposure to these data sources, indicating its ability to generalize across different synthesis conditions. These findings confirm that controlled fine-tuning effectively balances adaptation and generalization, while also highlighting the potential of incremental learning strategies to maintain robustness against evolving deepfake generation techniques.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Dwi Arman Prasetya,

Department Master of Information Technology,

University of Pembangunan Nasional Veteran Jawa Timur,

Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294.

Email: arman.prasetya.sada@upnjatim.ac.id

1. INTRODUCTION

Recent advances in Generative Artificial Intelligence (GenAI) have fundamentally transformed Text-to-Speech (TTS) and Voice Conversion (VC) technologies. The transition from traditional statistical approaches to deep generative models [1], followed by sequence-to-sequence architectures with attention mechanisms [2, 3], has substantially improved the naturalness and expressiveness of synthetic speech. More recently, progress in multi-speaker modeling, adversarial learning, and diffusion-based frameworks [4-8], has enabled high-fidelity voice synthesis and zero-shot voice cloning using minimal target data, further narrowing the perceptual gap between synthetic and authentic speech.

While these developments offer significant benefits, they also introduce serious security risks. Audio deepfakes have emerged as a critical threat to digital security, public trust, and social stability. Numerous studies report that AI-generated voice manipulation can be exploited for impersonation, financial fraud, misinformation campaigns, and attacks on voice-based systems such as biometric authentication and virtual assistants [9-11]. By enabling malicious actors to convincingly mimic specific individuals, audio deepfakes facilitate social engineering attacks and economic fraud, increasing the risk of financial loss and privacy violations [12]. Moreover, the growing prevalence of synthetic audio contributes to the erosion of trust in digital media, as distinguishing between bonafide and fabricated content becomes increasingly difficult [13, 14]. In political and societal domains, the misuse of audio deepfakes can further amplify disinformation and trigger social instability [15, 16], underscoring the urgent need for reliable detection methods and adaptive regulatory frameworks [9, 17].

Despite achieving high accuracy on benchmark datasets, many state-of-the-art audio deepfake detection systems exhibit limited generalization capability. Performance often degrades significantly when models are exposed to unseen data distributions, including novel synthesis techniques, diverse recording environments, or previously unobserved spoofing attacks [18-20]. Forensic analyses suggest that such detectors frequently overfit to dataset-specific rather than learning invariant manipulation cues [11, 18]. This generalization gap remains a major challenge, as models trained on constrained datasets struggle to maintain robustness under real-world deployment conditions characterized by acoustic variability and domain shifts [19, 20].

YAMNet has demonstrated strong robustness due to its large-scale pretraining on AudioSet and its efficient convolutional architecture. Prior studies show that YAMNet-based transfer learning provides stable performance under noisy and domain-shifted environments, including urban sound classification and acoustic scene recognition tasks [21, 22]. Additional evaluations highlight its adaptability in speech command recognition and embedded deployment scenarios, where classification reliability must be preserved under varying hardware and environmental constraints [23, 24]. These findings suggest that YAMNet embeddings capture high-level acoustic representations with strong generalization properties. However, processing frame level embeddings sequentially can introduce substantial computational overhead, particularly in resource-constrained or real-time security applications, which may limit scalability [25].

Building upon these observations, this study proposes an audio deepfake detection framework that leverages YAMNet as a high-level feature extractor combined with a lightweight Deep Neural Network (DNN) classifier. YAMNet is employed to extract frame level embedding vectors directly from raw audio waveforms, which are subsequently summarized using statistical aggregation operators, specifically the mean and standard deviation. This aggregation strategy captures both central tendency and variability of acoustic representations while significantly reducing feature dimensionality without introducing additional temporal modeling complexity or trainable parameters [24]. The resulting aggregated embeddings are then fed into a DNN to enable nonlinear discrimination between bonafide and spoofed audio. To explicitly evaluate generalization performance, the model is trained on publicly available datasets and assessed on a separate primary dataset containing previously unseen conditions. Through this cross-dataset evaluation, the proposed approach aims to provide a computationally efficient and robust defense against audio deepfakes while addressing the persistent generalization limitations of existing detection systems.

2. RESEARCH METHOD

To ensure that the issues explored in this research can be answered and explained chronologically through clear and accountable methods [26], this study adopts a systematic research methodology. The methodology follows a multi-stage process designed to ensure high model generalization across different audio domains. The process begins with data acquisition, utilizing both a public dataset for foundational training and a primary dataset for robustness validation, followed by feature extraction using YAMNet embeddings to capture complex acoustic patterns. The experimental workflow is then structured into two main phases: the optimization phase, which involves iterative DNN modeling and hyperparameter tuning on the public dataset to obtain optimal performance, and the adaptation phase, where the optimized model is evaluated

using the primary dataset. To address potential variations in data characteristics, a fine-tuning mechanism is incorporated to refine model parameters if the initial performance on the primary dataset is not satisfactory. Finally, a comprehensive evaluation is conducted to verify the effectiveness of the proposed approach, as illustrated in the detailed experimental flow presented in Figure 1.

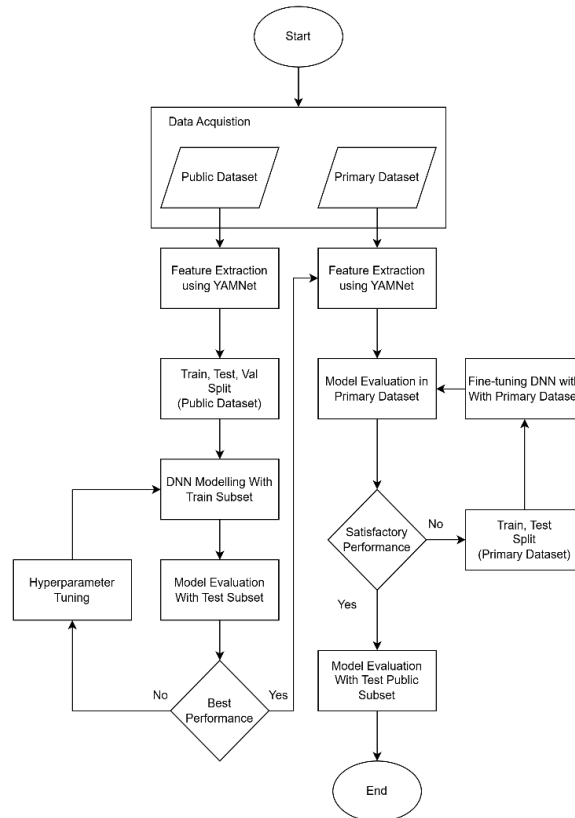


Figure 1. Research Workflow

2.1. Data Acquisition

This research utilizes the In the Wild secondary dataset [27], a collection of real world audio recordings featuring 54 different speakers across diverse environmental settings. The dataset comprises 19,963 bona-fide and 11,816 spoof samples sourced from public domains. All files were downsampled to 16 kHz to ensure uniformity and meet the specific input requirements of the YAMNet model.

To evaluate the generalization capability of the model, a primary dataset was developed involving 20 speakers who provided original voice recordings. Each speaker recorded 20 audio samples, consisting of 10 single-word utterances and 10 full-sentence utterances, resulting in a total of 400 bonafide audio samples. These recordings were subsequently duplicated using FishAudio mini [28] to generate high-fidelity synthetic voice clones, yielding an additional 400 spoofed audio samples. Consequently, the primary dataset comprises 800 audio samples in total, with a balanced distribution between bonafide and spoofed classes. By employing this primary dataset exclusively as an unseen test set, the study ensures that the model is rigorously evaluated under realistic conditions, including background noise, speaker variability, and diverse intonation patterns, thereby providing a robust and unbiased assessment of classification performance in practical deployment scenarios.

In addition, an external dataset obtained from [37] is used in this study to further evaluate the generalization capability of the proposed model under diverse synthesis conditions. This dataset consists of audio generated by various synthesis systems, including OpenAI, VALLE, and seedtts_files, and is utilized solely as supplementary evaluation data. It is important to note that this dataset is not involved in any training or fine-tuning process.

2.2. Feature Extraction using YAMNet

Feature extraction in this study is performed using YAMNet, a deep neural network pre trained on the AudioSet corpus. This model is designed to recognize 521 audio event classes and is particularly effective for capturing complex acoustic patterns in real world recordings. The architecture of YAMNet is based on MobileNetV1, which utilizes depthwise separable convolutions to achieve high computational efficiency without sacrificing classification accuracy. This efficiency makes it suitable for processing large scale datasets such as In the Wild and the high fidelity clones generated in this research.

The extraction process begins by resampling all audio signals to 16 kHz and converting them into log mel spectrograms. The audio is segmented into overlapping frames of 0.96 seconds with a 50% overlap, and each frame is processed by YAMNet to produce a 1024-dimensional embedding vector [25]. To obtain a fixed-length utterance level representation, frame level embeddings are aggregated using statistical pooling. Specifically, the mean operator captures the central tendency of the embeddings, while the standard deviation operator models temporal variability. The resulting vectors are concatenated to form a compact representation, which is subsequently used as input to the DNN classifier for bonafide and spoofed classification.

Table 1. Aggregation operators and output dimensions

Aggregation Operator	Output Dimension
Mean	1024
Standard Deviation	1024
Mean + Standard Deviation (Concatenate)	2048

2.3. Data Splitting

The In the Wild secondary dataset is partitioned into three subsets with a distribution of 75% for training, 15% for validation, and 10% for testing. This split results in 23,834 samples for training, while the 4,767 validation samples are used specifically to trigger the early stopping mechanism. The remaining 3,178 testing samples are utilized to evaluate model performance and select the best architecture to be further tested against the primary dataset for generalization. These partitioning employs stratified sampling to maintain a balanced ratio of bona fide and spoof audio across all sets. This type of dataset splitting strategy is often used to ensure that models can be trained effectively while preventing overfitting and enabling reliable evaluation on previously unseen data [29].

2.4. Classification Model Architecture

The classification phase of this research involves a comparative analysis of different Deep Neural Network (DNN) architectures to identify the most effective configuration for deepfake audio detection. Each model receives an utterance level feature vector extracted by YAMNet as its primary input. Two input feature dimensionalities are evaluated, 1024 and 2048. The architectures employ a progressively decreasing number of neurons to perform hierarchical feature compression, reducing the representation to a single output neuron.

To mitigate overfitting, Dropout layers are incorporated between dense layers, with a dropout rate of 0.5 applied after the first hidden layer and 0.2 applied to second hidden layer. All hidden layers utilize the Rectified Linear Unit (ReLU) activation function to introduce nonlinearity, while the final output layer employs a sigmoid activation function for binary classification between bonafide and spoofed audio. Model optimization is performed using the Adam optimizer with the Binary Crossentropy loss function. Additionally, an Early Stopping mechanism is applied with a patience of 10 epochs and best-weight restoration to enhance training stability and generalization.

Table 2. DNN architectures with 1024-dimensional input

Model	Total Hidden Layer	Neuron Distribution per Layer	Output Layer
Model A	3 Layers	1024, 512, 256	1 (Sigmoid)
Model B	4 Layers	1024, 512, 256, 128	1 (Sigmoid)
Model C	5 Layers	1024, 512, 256, 128, 64	1 (Sigmoid)

For higher-dimensional representations, additional DNN configurations are evaluated using a 2048-dimensional input feature vector. These models follow the same architectural design principles as the 1024-dimensional counterparts, employing a hierarchical reduction in neuron count and identical training strategies. This comparison enables an analysis of the impact of increased feature dimensionality on classification performance and generalization.

Table 3. DNN architectures with 2048-dimensional input

Model	Total Hidden Layer	Neuron Distribution per Layer	Output Layer
Model A	3 Layers	2048, 1024, 512	1 (Sigmoid)
Model B	4 Layers	2048, 1024, 512, 256	1 (Sigmoid)
Model C	5 Layers	2048, 1024, 512, 256, 128	1 (Sigmoid)

2.5. Experimental Setup

The experimental phase is designed to evaluate the ability of the proposed models to generalize effectively to new and unseen data. After the training process is completed using the secondary dataset, the best-performing model is selected and assessed under multiple evaluation scenarios to analyze both internal classification reliability and external generalization performance.

Three evaluation scenarios are considered in this study. In the first scenario, referred to as the primary full test, the trained model is directly evaluated on the entire primary dataset, which is treated as a completely unseen test set. This scenario aims to assess the intrinsic robustness of the model under realistic conditions without any prior exposure to target-domain data during training. This evaluation is also relevant in incremental learning frameworks, where models must preserve previously acquired knowledge while progressively adapting to new data distributions [30].

The second and third scenarios involve fine-tuning using different proportions of the primary dataset. In the second scenario, 20% of the primary data is allocated for fine-tuning while the remaining 80% is used for testing, representing a limited adaptation condition with scarce labeled target-domain data. In the third scenario, 40% of the primary dataset is used for fine-tuning and 60% for testing, reflecting a moderate adaptation condition with increased access to target-domain samples. These configurations are designed to simulate practical deployment situations in which the availability of labeled target-domain data varies, allowing the evaluation of model adaptability under different data constraints.

To analyze the impact of adaptation strategies, two fine-tuning approaches are evaluated. In the first approach, the model is fine-tuned without freezing any layers, allowing all network parameters to be updated. In the second approach, partial fine-tuning is applied by freezing selected early layers of the pretrained network, while the remaining layers are updated using the primary data. This general layer-freezing strategy is motivated by prior findings indicating that freezing early layers can help preserve transferable low-level representations and reduce overfitting when the amount of target-domain data is limited [31].

By comparing performance across the primary full test, full fine-tuning, and partial fine-tuning under different data split ratios, this experimental design enables a comprehensive analysis of the trade-off between adaptability and generalization. The results provide insights into how different fine-tuning strategies and target set sizes influence deepfake audio detection performance in realistic real-world scenarios.

2.6. Model Evaluation

To provide a comprehensive assessment [32], the models are evaluated using four key metrics that capture different aspects of classification success. Accuracy measures the overall frequency of correct predictions relative to the total dataset, calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

While high accuracy indicates strong overall performance, it can be misleading in cases of class imbalance. To address this, Precision is utilized to quantify the ability of the model to correctly identify positive instances out of all positive predictions, while Recall assesses the ability to recognize all actual positive data points. These are formulated as [33]:

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

Finally, because a trade off often exists between these two measures, the F1 score is employed as a harmonic mean to provide a balanced evaluation. Expressed as:

$$F1 - score = 2 \times \frac{recall \times precision}{recall + precision} \quad (3)$$

A high F1 score demonstrates that the model effectively produces correct positive predictions while successfully capturing all relevant positive instances, making it a critical metric for robust deepfake detection [34, 35].

3. RESULTS AND DISCUSSION

3.1. Public Dataset Modeling and Model Selection

This stage focuses on evaluating multiple Deep Neural Network (DNN) configurations trained on the public dataset with the objective of selecting the most reliable models for subsequent generalization experiments. A total of nine model configurations are examined by combining three DNN architectures with three different feature aggregation strategies: Mean, Standard Deviation (Std), and Mean + Std. Classification accuracy is employed as the primary criterion for model selection at this stage. The comparative results of all evaluated configurations, which form the basis for model selection, are presented in Table 4.

Table 4. Classification accuracy (%) of all model configurations on the public dataset

Feature Representation	Model	Accuracy (%)
Mean	Model A	99.06
Mean	Model B	99.06
Mean	Model C	98.87
Std	Model A	97.86
Std	Model B	98.11
Std	Model C	98.17
Mean + Std	Model A	99.24
Mean + Std	Model B	98.99
Mean + Std	Model C	98.93

From Table 4, the results reveal clear performance trends across both feature representations and model architectures. Feature representations based on Mean and Mean + Std consistently outperform those relying solely on Standard Deviation, indicating that central tendency information plays a more dominant role in discriminating between bonafide and spoofed audio than variance information alone. Although standard deviation captures temporal variability, it appears insufficient as a standalone descriptor for reliable classification in this task.

The performance differences between Model A, Model B, and Model C are relatively small, suggesting that increasing model depth does not significantly improve results. In some cases, deeper models provide only marginal gains or slight degradation, indicating that the YAMNet embeddings are already highly discriminative and do not require complex architectures.

Overall, the Mean + Std representation achieves the best performance, showing that combining complementary statistical features improves classification without adding complexity. Based on these results, the best-performing models are selected for further evaluation on the primary dataset to assess generalization on unseen data.

3.2. Full Test Evaluation on the Primary Dataset

This subsection evaluates the generalization performance of the selected models on the primary dataset, which is treated as a completely unseen test set. The primary dataset reflects realistic conditions with diverse speaker characteristics, intonation variations, and background noise, thereby providing a rigorous assessment of real-world deployment performance. No primary

data is used during training in this evaluation stage. The resulting classification performance of the selected models on the primary dataset is summarized in Table 5.

Table 5. Full test accuracy on the primary dataset

Model	Accuracy (%)
Mean + Std, Model A	62
Mean, Model A	62.37
Mean, Model B	61

As shown in Table 5, a substantial performance drop is observed compared to the results obtained on the public dataset, indicating a pronounced generalization gap between the source and target domains. Among the evaluated models, Mean, Model A achieves the highest accuracy on the primary dataset, although the margin over the other configurations remains relatively small.

To further contextualize these results, Table 6 presents a comparison with prior work [36], which also evaluates models under a direct testing setting using audio generated by the same synthesis system (FishAudio mini). Although the datasets differ in composition, both studies address the challenge of detecting spoofed audio generated from the same underlying model without prior exposure during training, making the comparison relevant for assessing robustness under consistent attack conditions.

Table 6. Performance comparison with prior work

Model	Correct Spoof	Correct Bonafide	Total Samples	Accuracy (%)
Wav2Vec 2.0	0 / 500	500/500	1000	50
Whisper	1 / 500	498/500	1000	49.9
AASIST	500 / 500	0 / 500	1000	50

Table 6 shows that the baseline models from [36] suffer from extreme class collapse, where predictions are dominated by a single class. As a result, the reported accuracy does not reflect meaningful classification performance, but rather indicates a failure to generalize under direct exposure to unseen generated audio.

To further assess the effectiveness of the proposed approach under the same challenge, Table 7 presents the detailed prediction results of the proposed models on the primary dataset.

Table 7. Detailed prediction results of the proposed models

Model	Correct Spoof	Correct Bonafide	Total Samples	Accuracy (%)
Mean + Std, Model A	106 / 400	390 / 400	800	62
Mean, Model A	100 / 400	399 / 400	800	62.37
Mean, Model b	90 / 400	398 / 400	800	61

In contrast to prior work, the proposed models demonstrate more stable and meaningful classification behavior. Although a bias toward bonafide predictions is still observed, the models are able to correctly identify a portion of spoof samples, indicating that the proposed approach avoids total class collapse and maintains partial discriminative capability across both classes. Furthermore, the results in Table 7 reveal a clear prediction imbalance, where spoof samples are more frequently misclassified as bonafide, suggesting that detecting synthesized audio remains the primary challenge under unseen conditions.

Overall, this evaluation highlights two key findings. First, direct cross-dataset generalization remains highly challenging. Second, compared to prior work, the proposed models achieve improved robustness and prediction stability, as evidenced by their ability to detect both classes without collapsing into a single dominant prediction. These findings motivate the need for adaptation strategies, which are explored in the subsequent subsection through fine-tuning using partial primary data.

3.3. Fine-Tuning with Limited Primary Data

The objective of this experiment is to investigate the effectiveness of fine-tuning without freezing any network layers when limited amounts of target-domain data are available. In this setting, all network parameters are allowed to update during training, enabling the model to fully adapt its learned representations to the characteristics of the primary dataset. This scenario reflects

a practical deployment condition where the model must rapidly adjust to new spoofing patterns, speaker variations, and acoustic environments, even though at the risk of degrading previously learned knowledge from the public dataset.

To simulate different levels of target-domain availability, two fine-tuning configurations are evaluated: a limited-data scenario and a moderate-data scenario. Model performance is assessed on both the primary dataset (ACC_P) to measure adaptation effectiveness and the secondary dataset (ACC_S) to evaluate retained generalization capability.

Table 8. Fine-tuning performance without layer freezing

Fine-Tuning Ratio	Model Configuration	ACC P (%)	ACC S (%)
20% Training 80% Testing	Mean + Std, Model A	94.84	88.17
	Mean, Model A	92.19	89.77
	Mean, Model B	89.53	91.79
40% Training 60% Testing	Mean + Std, Model A	94.79	83.79
	Mean, Model A	91.87	90.50
	Mean, Model B	92.92	90.59

As shown in Table 8, fine-tuning without layer freezing leads to a substantial improvement in primary dataset accuracy (ACC_P) across all evaluated model configurations. This confirms that unrestricted parameter updates enable the models to effectively adapt to the acoustic characteristics and spoofing present in the primary dataset. Among the evaluated configurations, Mean + Std, Model A consistently achieves the highest or near-highest ACC_P in both fine-tuning scenarios, indicating superior adaptation capability under limited and moderate data conditions.

In contrast, performance on the secondary dataset exhibits noticeable variation across models. While some configurations maintain relatively high ACC_S, others experience a clear performance drop as ACC_P increases. This divergence highlights the inherent trade-off between adaptation and generalization when fine-tuning is performed without freezing. Deeper or more flexible models appear more prone to source-domain performance degradation, whereas more compact architectures demonstrate better stability.

Overall, these results suggest that although fine-tuning without freezing is highly effective for improving target-domain performance, careful model selection is required to balance adaptation gains against generalization loss.

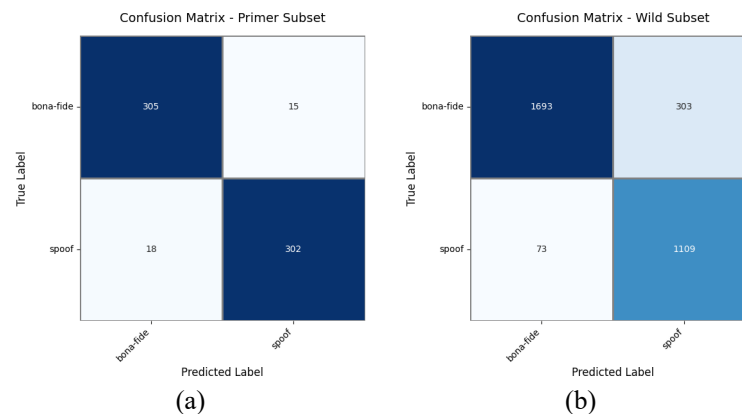


Figure 2. Confusion matrices of Mean + Std, Model A after 20% fine-tuning, (a) Primary Dataset, (b) Secondary Dataset

Figure 2 presents the confusion matrices of the best-performing configuration, Mean + Std, Model A, evaluated on both the primary and secondary datasets under the limited-data fine-tuning scenario, where 20% of the primary dataset is used for fine-tuning and 80% is reserved for testing. This setting represents the most challenging and realistic adaptation condition, as the model must adjust to target-domain characteristics with minimal labeled data.

On the primary dataset, the confusion matrix shows a substantial improvement in spoof detection performance, reflected by a high number of correctly classified spoof samples and a reduced false acceptance rate. This result indicates that even limited exposure to primary-domain

data enables the model to learn discriminative cues specific to the spoofing techniques present in the target domain.

On the secondary dataset, however, an increased number of bonafide samples are misclassified as spoof, revealing a noticeable shift in the decision boundary. This behavior suggests that the model becomes more conservative after fine-tuning, prioritizing spoof detection in the target domain at the expense of bonafide recall in the source domain. Such error patterns provide clear evidence of partial domain bias and highlight the trade-off between adaptation and generalization when fine-tuning is performed without layer freezing.

These findings indicate that while unrestricted fine-tuning effectively enhances target-domain spoof detection, it also introduces domain bias that degrades source-domain performance. This limitation motivates the adoption of a layer freezing strategy, which aims to preserve generalized acoustic representations while selectively adapting higher-level features, as explored in the following subsection.

3.4. Fine-Tuning with Layer Freezing Strategy

This subsection investigates the effectiveness of a layer freezing strategy during fine-tuning to address the generalization degradation observed in unrestricted fine-tuning. In this configuration, the early and intermediate layers of the network are frozen, while only the last two layers, specifically the layer before the output and the output layer itself, are kept trainable. The main objective of this strategy is to preserve the generalized acoustic representations learned from the public dataset, which are typically captured in the earlier layers, while enabling the final layers to adapt to the specific characteristics of the primary dataset. This approach is particularly suitable when only a small amount of labeled target-domain data is available, as it allows controlled adaptation to new spoofing patterns, speaker variations, and acoustic conditions, while reducing the risk of overfitting and maintaining stable performance across domains.

In this experiment, fine-tuning is performed by freezing the majority of the network layers and allowing only the final layers to be updated. As a result, low and mid-level acoustic features remain stable, while high-level decision boundaries are adjusted to better reflect the characteristics of the primary dataset. Two fine-tuning scenarios are evaluated to simulate different levels of target-domain data availability.

Table 9. Fine-tuning performance with layer freezing

Fine-Tuning Ratio	Model Configuration	ACC_P (%)	ACC_S (%)
20% Training 80% Testing	Mean + Std, Model A	92.81	91.41
	Mean, Model A	82.19	96.57
	Mean, Model B	76.41	98.49
40% Training 60% Testing	Mean + Std, Model A	92.71	92.13
	Mean, Model A	86.87	94.21
	Mean, Model B	81.25	97.48

As shown in Table 9, applying layer freezing leads to a more balanced performance across the primary and secondary datasets compared to unrestricted fine-tuning. Although a slight reduction in primary dataset accuracy (ACC_P) is observed, the models exhibit a clear improvement in secondary dataset accuracy (ACC_S), indicating enhanced preservation of source-domain knowledge.

Among the evaluated configurations, Mean + Std, Model A consistently achieves the highest ACC_P under both fine-tuning scenarios, confirming its robustness in adapting to the primary dataset. Meanwhile, models based on the Mean representation show superior ACC_S, suggesting that freezing layers effectively stabilizes generalized acoustic features and limits excessive adaptation. These results demonstrate that the layer freezing strategy successfully constrains parameter updates, preventing over-specialization to the target domain while maintaining meaningful adaptation capability.

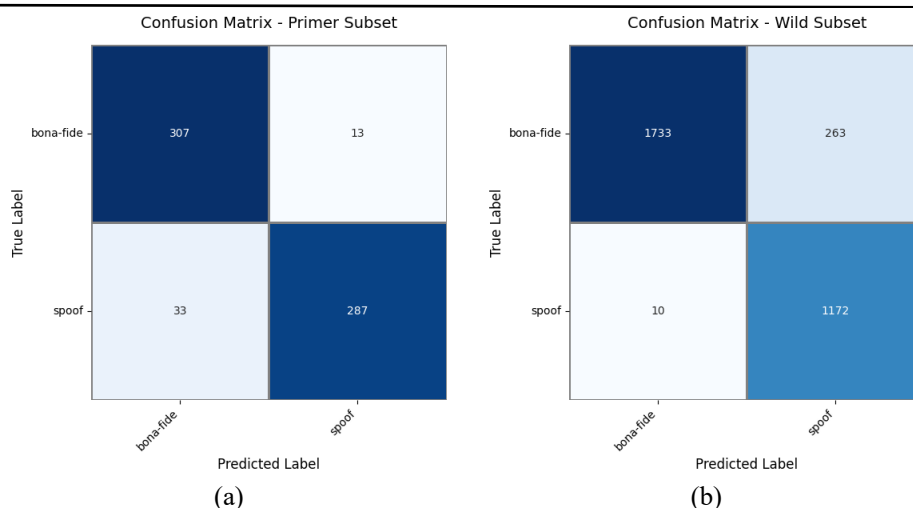


Figure 3. Confusion matrices of Mean + Std, Model A after 20% fine-tuning, (a) Primary Dataset, (b) Secondary Dataset

Figure 3 presents the confusion matrices of the best-performing configuration, Mean + Std, Model A, evaluated under the limited-data fine-tuning scenario with layer freezing. On the primary dataset, the confusion matrix indicates a modest increase in misclassified spoof samples compared to unrestricted fine-tuning. However, the overall classification performance remains stable, showing that essential discriminative cues are retained despite constrained adaptation.

On the secondary dataset, the confusion matrix reveals a noticeable reduction in bonafide misclassification. This improvement confirms that layer freezing effectively mitigates the decision boundary shift observed in the no-freeze setting and enhances generalization stability.

The observed error patterns suggest that layer freezing provides an effective compromise between adaptation and generalization. While unrestricted fine-tuning prioritizes target-domain performance at the cost of source-domain degradation, the freezing strategy preserves core acoustic representations and reduces catastrophic forgetting. Consequently, this approach offers a more stable solution for real-world deployment scenarios where robustness across domains is essential.

3.5. Best Model Evaluation

This subsection presents a detailed evaluation of the selected best-performing model, Mean + Std, Model A with layer freezing, under the 20–80 fine-tuning scenario, where 20% of the primary dataset is used for adaptation and 80% is reserved for testing. This configuration represents the most realistic deployment condition, in which only limited labeled target-domain data are available. Model performance is evaluated using precision, recall, and F1-score on both the primary and secondary datasets. The analysis emphasizes class-wise performance to provide deeper insight into the model's behavior in distinguishing between bona-fide and spoofed audio under different domain conditions.

Table 10. Classification report on the primary dataset

Class	Precision	Recall	F1-score	Support
Bona-fide	0.90	0.96	0.93	320
Spoof	0.96	0.90	0.93	320
Macro Avg	0.93	0.93	0.93	640
Weighted Avg	0.93	0.93	0.93	640

On the primary dataset, the model achieves balanced performance across both bona-fide and spoof classes, as reflected by identical F1-scores of 0.93. The high recall for bona-fide samples indicates that the model effectively minimizes false rejections of genuine speech, which is crucial for maintaining usability in real-world applications. At the same time, the spoof class exhibits high precision, suggesting that false acceptance of spoofed audio is effectively controlled. This balance indicates that the integration of YAMNet-based embeddings with statistical aggregation and

controlled fine-tuning enables the model to learn discriminative features that generalize well to unseen spoofing methods present in the primary dataset.

Table 11. Classification report on the secondary dataset

Class	Precision	Recall	F1-score	Support
Bona-fide	0.99	0.87	0.93	1996
Spoof	0.82	0.99	0.90	1182
Macro Avg	0.91	0.93	0.91	3178
Weighted Avg	0.93	0.91	0.92	3178

On the secondary dataset, the model demonstrates a distinct yet interpretable performance shift. The very high recall for the spoof class (0.99) indicates that the model remains highly sensitive to spoofing learned during adaptation. However, this sensitivity is accompanied by a reduction in bona-fide recall, reflecting a more conservative decision strategy. In practice, this behavior implies that the model prioritizes security by aggressively detecting spoofed audio, even at the cost of increased false rejections of bonafide samples.

Despite this trade-off, the macro-averaged metrics remain stable, confirming that overall performance does not collapse under domain shift. This observation suggests that the layer freezing strategy successfully mitigates catastrophic forgetting by preserving core acoustic representations learned from the public dataset while allowing adaptation to new domains.

To further evaluate the robustness of the selected model, an additional experiment is conducted on external datasets obtained from [37], which are not part of the primary or secondary datasets used in the previous experiments. These datasets consist of audio generated from different synthesis systems, namely OpenAI, VALLE, and seedtts_files, representing diverse synthesis characteristics and varying levels of domain shift. Importantly, the fine-tuning process is performed exclusively using the primary dataset, without any exposure to the external datasets during training or adaptation, ensuring that the evaluation strictly reflects the model's ability to generalize to unseen data distributions.

Table 12. Performance improvement after fine-tuning across external datasets

Model	Accuracy Before (%)	Accuracy After (%)	Improvement (%)
OpenAI	83.16	87.66	+ 4.50
VALLE	43.15	78.94	+ 35.79
seedtts_files	33.88	64.6	+ 30.72

As shown in Table 12, fine-tuning consistently improves performance across all external datasets. The most substantial improvements are observed in VALLE and seedtts_files, indicating that the model initially struggles with these datasets due to stronger domain shifts. After adaptation, the model significantly recovers its performance, demonstrating improved robustness even under substantial domain shifts.

In contrast, the improvement on the OpenAI-generated dataset is relatively smaller, suggesting that the model already captures relevant characteristics of this data prior to fine-tuning. These results indicate that the effectiveness of fine-tuning is strongly influenced by the degree of domain mismatch between the source and target data distributions.

Overall, the evaluation across primary, secondary, and external datasets highlights a fundamental trade-off between adaptation and generalization in cross-domain deepfake detection. The results demonstrate that fine-tuning significantly enhances the model's ability to adapt to new data distributions, even when only limited target-domain data are available. Importantly, these improvements are achieved without direct exposure to fully unseen datasets during training, indicating that the model learns transferable representations that extend beyond the observed domains. The consistent performance gains across different synthesis conditions further emphasize the effectiveness of fine-tuning as a lightweight yet powerful adaptation strategy.

However, the presence of domain shifts and continuously evolving spoofing techniques suggests that static fine-tuning alone may not be sufficient for long-term robustness. In this context, incremental learning emerges as a critical direction, enabling the model to continuously incorporate new data and adapt to emerging attack patterns without degrading previously learned knowledge. By leveraging incremental updates, the proposed framework can be extended to maintain its

effectiveness over time, making it more suitable for real-world deployment where the characteristics of synthesized audio are constantly evolving.

3.6. Error Analysis

To further analyze the behavior of the proposed model, an error analysis is conducted by examining the distribution of embedding representations and their classification outcomes. This analysis aims to understand how well the model generalizes to new datasets and to identify patterns associated with misclassification, particularly during the transition from secondary data to primary data.

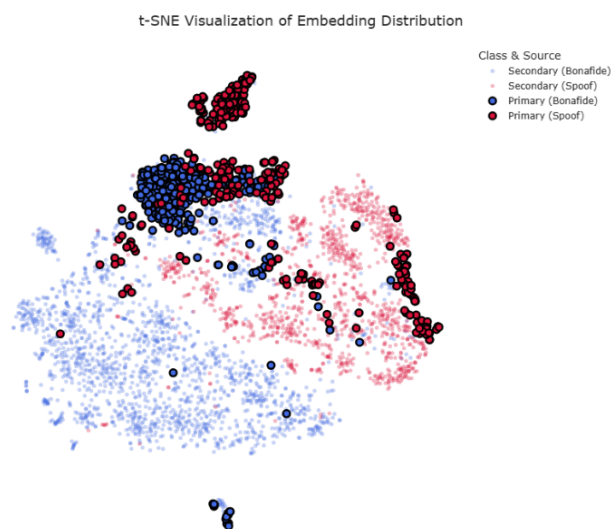


Figure 4. t-SNE visualization of embedding distribution for primary and secondary datasets

Figure 4 shows the distribution of embeddings from both the primary and secondary datasets. It can be observed that samples from the primary dataset follow a similar structure to those from the secondary dataset, forming consistent clusters for bonafide and spoof. This indicates that the pretrained YAMNet embeddings are able to generalize effectively to new data. The clear separation between bonafide and spoof suggests that the feature extraction process alone is already strong in distinguishing between the two classes. In particular, the aggregation of frame-level embeddings using mean and standard deviation is sufficient to produce discriminative representations, even before applying more complex classification models.

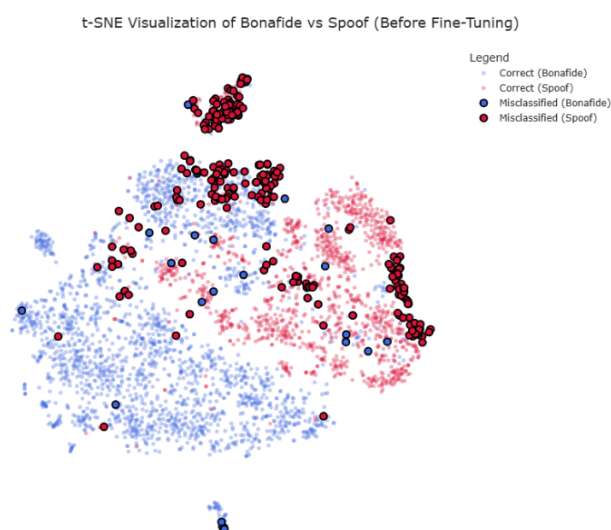


Figure 5. t-SNE visualization of classification results before fine-tuning

Figure 5 shows the classification results prior to the fine-tuning process. Although the overall separation between bonafide and spoof is already visible, a considerable number of misclassified samples can still be observed. This indicates that the model trained on the secondary dataset has not yet fully adapted to the primary dataset. Notably, several spoof samples are positioned close to the bonafide cluster learned from the secondary data, causing these spoof samples to be classified as bonafide. This behavior suggests that, during the transition to the primary dataset, the model still relies heavily on the bonafide patterns learned from the secondary dataset. As a result, spoof samples that exhibit similar characteristics to bonafide are more likely to be misclassified, leading to a higher number of errors in the spoof class.

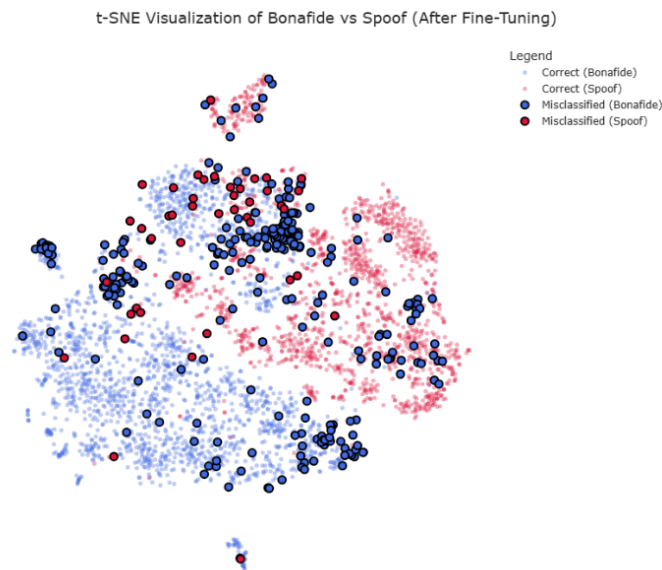


Figure 6. t-SNE visualization of classification results after fine-tuning

Figure 6 shows the classification results after the fine-tuning process. Compared to the previous stage, the separation between bonafide and spoof becomes more refined, and the number of misclassified samples is reduced. This indicates that fine-tuning enables the model to better adapt its decision boundaries to the characteristics of the primary dataset. Nevertheless, some misclassification still persists, particularly in cases where spoof samples closely resemble bonafide.

Overall, this analysis confirms that the feature extraction approach based on YAMNet embeddings provides strong discriminative capability in separating bonafide and spoof. The fine-tuning process further enhances the model's ability to adapt from secondary data to primary data, even when using a relatively simple DNN model. The remaining errors are primarily associated with spoof samples that closely resemble bonafide, highlighting the inherent challenge of distinguishing between the two classes in such conditions.

3.7. Limitations

This study has several limitations related to the evaluation setup and feature extraction strategy. The experimental validation is conducted on the secondary and primary datasets, while additional external datasets are used solely for cross-domain evaluation. Although this design enables the analysis of cross-domain performance and the effectiveness of fine-tuning strategies, the evaluation scenario remains limited in terms of dataset diversity.

Another limitation concerns the feature extraction approach adopted in this study. The proposed method relies solely on pretrained YAMNet embeddings as the primary feature representation, without direct comparison to other feature extraction techniques such as CQCC, MFCC, LFCC, or alternative deep learning-based extractors. While prior studies have shown that these alternative methods can achieve strong performance, they often introduce higher computational complexity and potential bias due to manual feature engineering [25]. In

comparison, this study utilizes a pretrained YAMNet model to automatically extract features from audio signals without human intervention, resulting in a more efficient pipeline.

In addition, the aggregation of frame-level embeddings using statistical operations is specifically designed to reduce the computational cost associated with processing high-dimensional and variable-length frame-level outputs from YAMNet. By converting these frame-level representations into a single fixed-length vector, the approach simplifies the input structure and enables the use of a lightweight DNN model.

4. CONCLUSION

This study presents a deepfake audio detection framework that integrates pretrained YAMNet embeddings with statistical feature aggregation and Deep Neural Network classifiers. Experiments on the public dataset demonstrate that Mean and Mean + Std aggregation strategies consistently outperform Standard Deviation-based features, while increasing network depth provides only marginal performance gains. Cross-dataset evaluation on a newly collected primary dataset reveals a pronounced generalization gap, indicating that direct deployment without adaptation is insufficient under realistic acoustic and spoofing conditions.

Fine-tuning experiments show that adaptation with limited target-domain data consistently and significantly improves detection performance. Under the 20–80 fine-tuning scenario, the best-performing model employing a layer freezing strategy achieves an accuracy of 0.93 on the primary dataset and 0.91 on the secondary dataset. Further analysis using precision, recall, and F1-score confirms balanced class-wise performance on the primary dataset, with macro-averaged metrics reaching 0.93, while stable generalization is maintained on the secondary dataset with a macro-averaged F1-score of 0.91.

Additional evaluation on external datasets further demonstrates that the proposed model benefits from fine-tuning even without prior exposure to these data sources, indicating its ability to generalize across different synthesis conditions. These findings validate that controlled fine-tuning with layer freezing offers an effective trade-off between adaptation and generalization for real-world deepfake audio detection. Furthermore, the results highlight the potential of incremental learning strategies to continuously improve model robustness as new spoofing techniques emerge.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The data associated with this study are publicly available online in the replication package. [\[URL\]](#)

AUTHOR CONTRIBUTIONS

Hakam Dzakwan Diash1: Conceptualization; Data Curation; Data Labeling; Methodology; Software; Validation; Writing–Original Draft; Visualization. **Dwi Arman Prasetya:** Supervision; Validation; Writing–Review & Editing. **Alfan Rizaldy Pratama:** Supervision; Validation; Writing–Review & Editing. **Tresna Maulana Fahrudin:** Supervision; Validation; Writing–Review & Editing.

REFERENCES

- [1] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” Sep. 19, 2016, *arXiv*: arXiv:1609.03499. doi: 10.48550/arXiv.1609.03499.
- [2] Y. Wang *et al.*, “Tacotron: Towards End-to-End Speech Synthesis,” in *Interspeech 2017*, ISCA, Aug. 2017, pp. 4006–4010. doi: 10.21437/Interspeech.2017-1452.
- [3] Y. Ren *et al.*, “FastSpeech: Fast, Robust and Controllable Text to Speech”.
- [4] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, and F. Ren, “Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis”.

- [5] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome: IEEE, Sep. 2018, pp. 2100–2104. doi: 10.23919/EUSIPCO.2018.8553236.
- [6] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece: IEEE, Dec. 2018, pp. 266–273. doi: 10.1109/SLT.2018.8639535.
- [7] S. Liu, Y. Cao, D. Su, and H. Meng, "DiffSVC: A Diffusion Probabilistic Model for Singing Voice Conversion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia: IEEE, Dec. 2021, pp. 741–748. doi: 10.1109/ASRU51503.2021.9688219.
- [8] C. Wang *et al.*, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," Jan. 05, 2023, *arXiv*: arXiv:2301.02111. doi: 10.48550/arXiv.2301.02111.
- [9] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, pp. 144497–144529, 2023, doi: 10.1109/ACCESS.2023.3344653.
- [10] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," *Algorithms*, vol. 15, no. 5, p. 155, May 2022, doi: 10.3390/a15050155.
- [11] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, no. 7, p. 1989, Mar. 2025, doi: 10.3390/s25071989.
- [12] F. Romero-Moreno, "Deepfake detection in generative AI: A legal framework proposal to protect human rights," *Comput. Law Secur. Rev.*, vol. 58, p. 106162, Sep. 2025, doi: 10.1016/j.clsr.2025.106162.
- [13] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," *IEEE Access*, vol. 11, pp. 132652–132682, 2023, doi: 10.1109/ACCESS.2023.3333866.
- [14] S. Kopecky, "Challenges of Deepfakes," in *Intelligent Computing*, vol. 1016, K. Arai, Ed., in *Lecture Notes in Networks and Systems*, vol. 1016, Cham: Springer Nature Switzerland, 2024, pp. 158–166. doi: 10.1007/978-3-031-62281-6_11.
- [15] Dr. B. Dhiman, "Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century," *Qeios*, Dec. 2023, doi: 10.32388/DPLE2L.
- [16] A. Hery, O. Joseph, O. Femi, and H. Luz, "Audio deepfakes: threats to voice assistants and voice-activated systems".
- [17] A. O. Vaidya, M. Dangore, V. K. Borate, N. Raut, Y. K. Mali, and A. Chaudhari, "Deep Fake Detection for Preventing Audio and Video Frauds Using Advanced Deep Learning Techniques," in *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Kothamangalam, Kerala, India: IEEE, May 2024, pp. 1–6. doi: 10.1109/RAICS61201.2024.10689785.
- [18] I. Amerini *et al.*, "Deepfake Media Forensics: Status and Future Challenges," *J. Imaging*, vol. 11, no. 3, p. 73, Feb. 2025, doi: 10.3390/jimaging11030073.
- [19] X. Li, P.-Y. Chen, and W. Wei, "Measuring the Robustness of Audio Deepfake Detectors," Mar. 21, 2025, *arXiv*: arXiv:2503.17577. doi: 10.48550/arXiv.2503.17577.
- [20] X. Zhang, J. Yi, and J. Tao, "Towards Robust Audio Deepfake Detection: A Evolving Benchmark for Continual Learning," Aug. 13, 2024, *arXiv*: arXiv:2405.08596. doi: 10.48550/arXiv.2405.08596.
- [21] L. Liu *et al.*, "YAMNet-based transfer learning for compact noise classification in urban and wireless systems," *EURASIP J. Wirel. Commun. Netw.*, vol. 2025, no. 1, p. 74, Sep. 2025, doi: 10.1186/s13638-025-02483-8.
- [22] H. Zhong, J. M. Buchholz, J. Maclaren, S. Carlile, and R. Lyon, "A dataset and model for auditory scene recognition for hearing devices: AHEAD-DS and OpenYAMNet," Jan. 13, 2026, *arXiv*: arXiv:2508.10360. doi: 10.48550/arXiv.2508.10360.
- [23] S. Lachenani, H. Kheddar, and M. Ouldzmirli, "Improving Pretrained YAMNet for Enhanced Speech Command Detection via Transfer Learning," in *2024 International Conference on*

- Telecommunications and Intelligent Systems (ICTIS)*, Djelfa, Algeria: IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/ICTIS62692.2024.10894266.
- [24] M. Lamrini, M. Y. Chkouri, and A. Touhafi, “Evaluating the Performance of Pre-Trained Convolutional Neural Network for Audio Classification on Embedded Systems for Anomaly Detection in Smart Cities,” *Sensors*, vol. 23, no. 13, p. 6227, Jul. 2023, doi: 10.3390/s23136227.
- [25] K. Shahzad, S. Farhan, Yasin-UI-Haq, R. Sana, and S. Pathan, “Enhancing Voice Spoofing Detection: A Hybrid Approach With VGGish-LSTM Model for Improved Security in Automatic Speaker Verification Systems,” *IEEE Access*, vol. 13, pp. 40682–40702, 2025, doi: 10.1109/ACCESS.2025.3544639.
- [26] D. A. Prasetya, A. Sanusi, G. Chandrarin, E. Roikhah, I. Mujahidin, and R. Arifuddin, “Small and Medium Enterprises Problem and Potential Solutions for Waste Management,” *J. Southwest Jiaotong Univ.*, vol. 54, no. 6, p. 21, 2019, doi: 10.35741/issn.0258-2724.54.6.21.
- [27] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, “Does Audio Deepfake Detection Generalize?,” in *Interspeech 2022*, ISCA, Sep. 2022, pp. 2783–2787. doi: 10.21437/Interspeech.2022-108.
- [28] S. Liao *et al.*, “Fish-Speech: Leveraging Large Language Models for Advanced Multilingual Text-to-Speech Synthesis,” Nov. 09, 2024, *arXiv*: arXiv:2411.01156. doi: 10.48550/arXiv.2411.01156.
- [29] A. Muhaimin, W. Wibowo, and P. A. Riyantoko, “Multi-label Classification Using Vector Generalized Additive Model via Cross-Validation,” *J. Inf. Commun. Technol.*, vol. 22, 2023, doi: 10.32890/jict2023.22.4.5.
- [30] A. F. Adziima, A. Muhaimin, A. T. Damaliana, A. R. Pratama, K. M. Hindrayani, and D. A. Prasetya, “Incremental Learning for Drought Prediction with CatBoost on Time-Series Data,” in *2025 IEEE 11th Information Technology International Seminar (ITIS)*, Mataram, Indonesia: IEEE, Oct. 2025, pp. 400–405. doi: 10.1109/ITIS67966.2025.11308862.
- [31] D. Soekhoe, P. Van Der Putten, and A. Plaat, “On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks,” in *Advances in Intelligent Data Analysis XV*, vol. 9897, H. Boström, A. Knobbe, C. Soares, and P. Papapetrou, Eds., in *Lecture Notes in Computer Science*, vol. 9897, Cham: Springer International Publishing, 2016, pp. 50–60. doi: 10.1007/978-3-319-46349-0_5.
- [32] D. A. Prasetya, A. P. Sari, P. A. Riyantoko, and T. M. Fahrudin, “The Effect of Information Quality and Service Quality on User Satisfaction of the Government of Kabupaten Malang,” *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 32–42, Jun. 2023, doi: 10.38043/tiers.v4i1.4328.
- [33] M. Nasrudin, S. S. M. Wara, A. Muhaimin, N. I. Nirmalasari, and M. R. Arfiana, “Detection of Ventricular Septal Defect in Pediatric Cardiac Ultrasound Videos Using Parasternal View and Faster R-CNN,” vol. 15, no. 1, 2026.
- [34] T. M. Fahrudin, I. Syarif, and A. R. Barakbah, “Data Mining Approach for Breast Cancer Patient Recovery,” *Emit. Int. J. Eng. Technol.*, vol. 5, no. 1, pp. 36–71, Jul. 2017, doi: 10.24003/emitter.v5i1.190.
- [35] M. Idhom, D. A. Prasetya, P. A. Riyantoko, T. M. Fahrudin, and A. P. Sari, “Pneumonia Classification Utilizing VGG-16 Architecture and Convolutional Neural Network Algorithm for Imbalanced Datasets,” *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 73–82, Jun. 2023, doi: 10.38043/tiers.v4i1.4380.
- [36] E. Issa, “AFAD-MSA: Dataset and Models for Arabic Fake Audio Detection,” *Computation*, vol. 14, no. 1, p. 20, Jan. 2026, doi: 10.3390/computation14010020.
- [37] M. KOKLU, B. Isgor, and C. Tumer, “Hybrid Feature-Based Two-Stage Framework for Audio Deepfake Detection and Generative Model Attribution,” *Adv. Eng. Intell. Syst.*, vol. 005, no. 01, Mar. 2026, doi: 10.22034/aeis.2026.566863.1400.