Serverless Data Pipelines for bird sound classification in cloud environments

1.Introduction

The rapid advancements in serverless cloud computing and artificial intelligence (AI) have created new opportunities for real-time sound classification across various domains, including wildlife monitoring, environmental sound analysis, security surveillance, and healthcare applications (Feng et al., 2024). Traditional sound classification methods often rely on centralized cloud computing, where raw audio data is transmitted to remote servers for processing. While these cloud-based models offer high computational power, they come with significant drawbacks, such as high latency, increased bandwidth consumption, and a dependence on stable internet connectivity. Moreover, many conventional approaches involve manual observation and analysis, in which human experts listen to audio recordings and annotate them by hand. This process is not only time-consuming but also susceptible to human error and scalability limitations.

Acoustic monitoring supports health, environmental, and biodiversity assessment, including heart sound analysis for health detection (Dong et al., 2013; Wimmer et al., 2010), bird sound analysis for habitat monitoring (Höchst et al., 2022), and environmental sound recognition for surveillance and forest conservation (Bandara et al., 2023). Bird sound classification is essential for tracking species distribution, assessing habitat health, and understanding behavioral patterns (Höchst et al., 2022). Acoustic monitoring aids biodiversity research by enabling automated species identification and long-term ecological surveillance.

Bird sound classification plays a vital role in ecological research, conservation efforts, and biodiversity monitoring. Automated acoustic monitoring allows researchers to track species distribution, assess habitat health, and study behavioral patterns over time (Xie et al., 2022). Traditional methods, which often rely on manual listening and annotation, are time-consuming and prone to human error (Höchst et al., 2022). Leveraging cloud-based AI systems for bird sound classification enhances efficiency and enables real-time data processing at scale.

Swaminathan et al. (2024) demonstrated that Wav2Vec transfer learning outperforms traditional models in bird sound classification, achieving an F1-score of 0.89 using the Xeno-Canto dataset with 10 bird species and 100 samples per species. Compared to CNN-LSTM and Transformer-based models, Wav2Vec's ability to process raw waveforms proved more effective, highlighting its potential for real-time bird sound classification in cloud-based AI systems.

Building on the findings of Swaminathan et al. (2024), which demonstrated that Wav2Vec transfer learning outperforms traditional models in bird sound classification, this study implements a serverless data pipeline framework for real-time bird sound classification in cloud environments**.** Recent studies, such as Shojaee Rad and Ghobaei-Arani (2024), have reviewed serverless computing strategies and emphasized their suitability for scalable, cost-efficient, and event-driven machine learning workflows. Similarly, Mbata et al. (2024) categorized modern serverless pipeline tools and highlighted their growing adoption in real-time data engineering tasks, including sound classification. Using a dataset from Xeno-Canto, this research focuses on three bird species commonly found in Thammasat University to evaluate the effectiveness of batch processing models. The proposed framework is designed to optimize latency and accuracy trade-offs through a six-step data pipeline implemented on

the AWS cloud platform, consisting of (1) Data ingestion, (2) Preprocessing, (3) Preprocessing orchestration, (4) Model training, (5) Model deployment, and (6) Inference & Prediction. By leveraging serverless computing, this study aims to develop a scalable and efficient cloud-based classification pipeline, contributing to advancements in automated avian acoustic monitoring and real-time biodiversity research.

Research objective

This study examines a serverless data pipeline framework for real-time bird sound classification in cloud environments, focusing on the batch processing model. Using a dataset from Xeno-Canto, featuring three bird species commonly found in Thammasat University.

2.Dataset Description

The experiments were conducted using bird sound recordings collected from the Xeno-Canto online repository. These recordings focused on three bird species that are commonly found in and around the Thammasat University campus: the White-breasted Waterhen, the Black-browed Reed Warbler, and the Stork-billed Kingfisher. Each audio file was preprocessed to reduce background noise and converted into a uniform format—16-bit mono WAV files sampled at 16 kHz—to ensure consistency across the dataset. The recordings were then divided into training, validation, and testing subsets. For instance, the White-breasted Waterhen included 182 samples for training, 20 for validation, and 25 for testing. The diversity in audio content, such as calls, songs, and alarm calls, provided a rich dataset for training a robust classification model.

The audio recordings of the three bird species used in this experiment were sourced from the Xeno-Canto website. The number of audio files for each species is presented in Table 2

Table 2 Number of registration of species

Species	Train	Validation	Test
White-breasted Waterhen	182	20	25
Black-browed Reed Warbler	164	18	14
Stork-billed Kingfisher	152	16	16

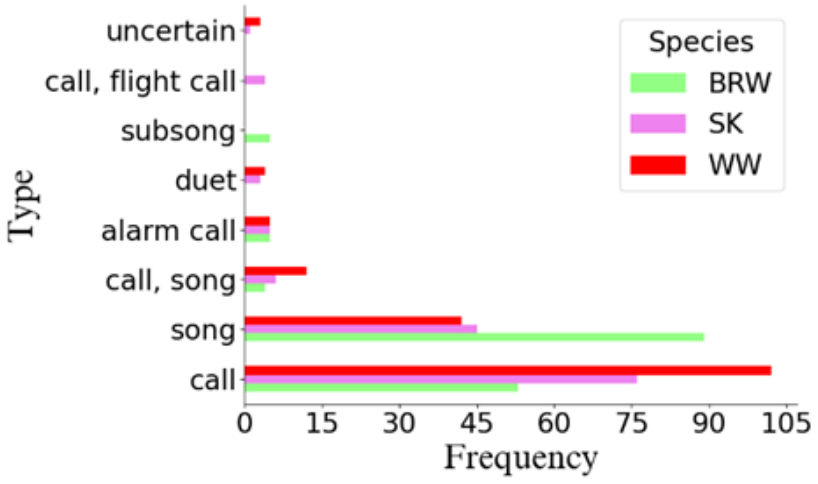

Figure 2 Training Dataset: 8 Sample Sound Types

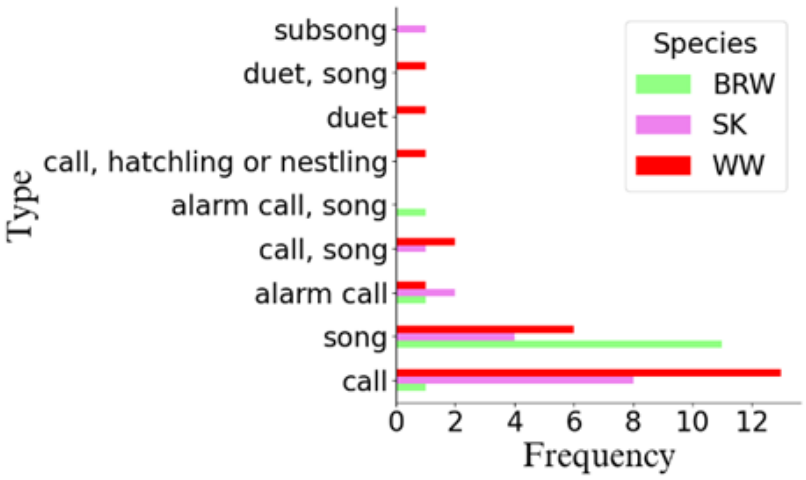

Figure 3 Testing Dataset: Types

Table 3 Training Dataset: Sound Type Distribution

Type	WW	BRW	SK	Total
call	102	53	76	231
song	42	89	45	176
call, song	12	4	6	22
alarm call	5	5	5	15
duet	4	0	3	7
subsong	0	5	0	5
uncertain	3	0	1	4
call, subsong	0	4	0	4
call, flight call	0	0	4	4
aberrant, call	1	0	2	3
alarm call, call	2	0	1	3
flight call	0	0	3	3
song, subsong	0	2	0	2
?	0	1	1	2
-	0	0	2	2
Low-intensity song.	1	0	0	1
call, maybe a call	1	0	0	1
call, duet	1	0	0	1
call, chorus	1	0	0	1
duet, flight call	0	0	1	1
duet, song	1	0	0	1
call, ?	0	0	1	1
nocturnal flight call	1	0	0	1
bird	1	0	0	1
song, cackling song	1	0	0	1
song, rattling	1	0	0	1
alarm call, contact & alarm calls	1	0	0	1
aberrant, call, duet, song	1	0	0	1

Type	WW	BRW	SK	Total
territorial call??	0	0	1	1
call, single note calls	0	1	0	1

Table 2 presents the distribution of different sound types in the training dataset for three bird species: WW, BRW and SK. “Call” is the most common sound type across all three species, with 102, 53, and 76 records for WW, BRW, and SK, respectively. “Song” is also frequently recorded, particularly for the BRW (89 records). Other sound types such as “alarm call,” “duet,” and “subsong” appear but are far less frequent. Some rare sound types include “nocturnal flight call” and “territorial call??”, appearing only once. A few sound types were labeled with uncertainty, as seen with "call, ?" and "uncertain." Figure 4 provides a visual representation of the sound types in the training dataset, focusing on eight key categories. These categories likely include the most frequently occurring sound types mentioned in Table 2. Figure 5 illustrates the distribution of sound types in the testing dataset, which consists of audio recordings reserved for evaluating model performance.

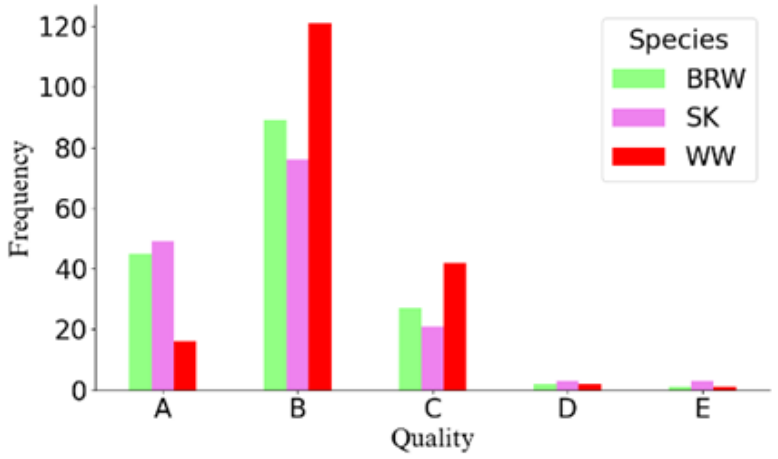

Figure 4 Training Dataset: Quality

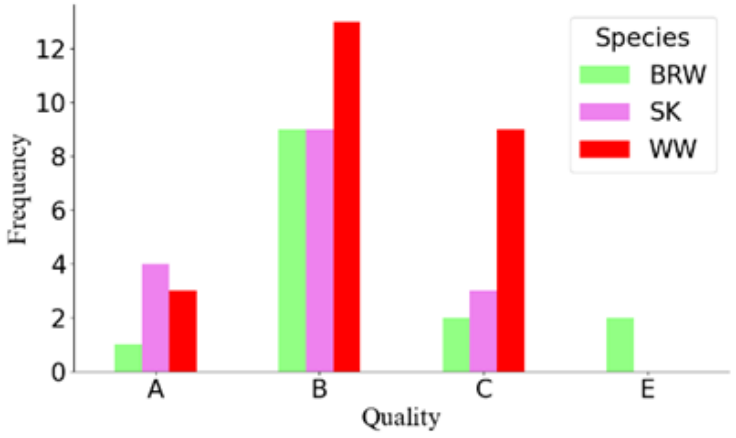

Figure 5 Testing Dataset: Quality

As shown in Figure 4 the training dataset consists of audio qualities A, B, C, D, and E, while the testing dataset includes audio qualities A, B, and C. The quality E is present only for the Black-browed Reed Warbler species, with no data for this quality from other species. Additionally, quality D does not have data for any species as shown in Figure 5

3. 5V analysis

Value

Bird sound classification delivers value in biodiversity monitoring, species identification, and ecological research. By automating the recognition of bird vocalizations, the project contributes to conservation efforts, habitat assessment, and long-term ecological surveillance. The insights generated from this system can support decision-making by researchers, conservationists, and policy-makers.

Veracity

The dataset may contain mislabeled samples, background noise, or inconsistent recording conditions. This variability introduces uncertainty, which can affect model performance. To address this, the pipeline includes a preprocessing stage powered by AWS Lambda that performs noise reduction, normalization, and segmentation, helping to improve the accuracy and reliability of the input data before training and inference.

Variety

The project deals with unstructured audio recordings alongside structured metadata such as species labels, location, and sound types (e.g., call, song, alarm call). This diversity requires a flexible and modular data processing pipeline capable of handling heterogeneous data formats and enriching raw audio with meaningful annotations for supervised learning.

Velocity

Although the initial ingestion of audio files is performed in batch mode (manual upload to Amazon S3), the system is designed for real-time inference after deployment.

Volume

Audio recordings, particularly high-resolution .wav files, are storage-intensive. As the dataset expands to include more species, vocalization types, and geographic areas, the data volume increases significantly. By leveraging scalable AWS services such as Amazon S3 and serverless compute, the architecture is well-suited to handle the increasing volume without sacrificing performance.

4. Data pipeline implementation and data pipeline architecture

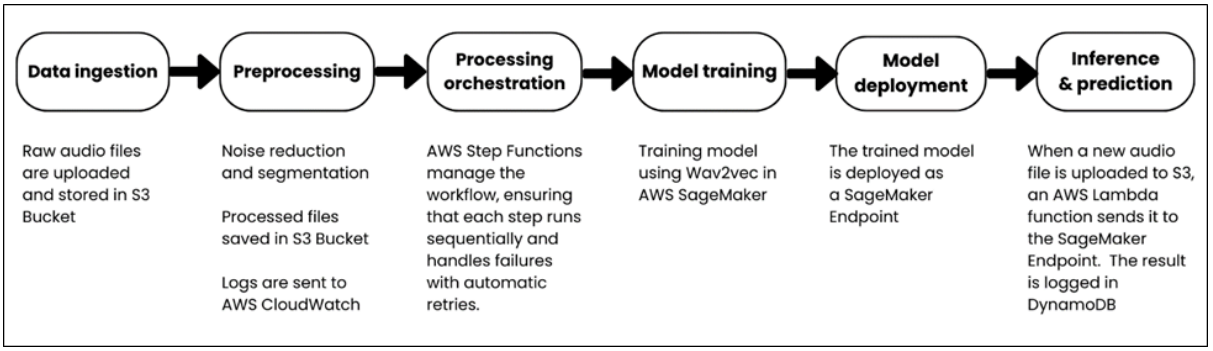

Figure 1 Proposed framework for data pipelines for real-time bird sound classification in cloud system

Proposed framework

This study presents a serverless data pipeline framework for real-time bird sound classification using AWS cloud services and a deep learning model. The framework is designed to automate the end-to-end processing of bird vocalization data, enabling scalable, event-driven execution without the need for dedicated server infrastructure. The pipeline comprises six sequential stages that handle data ingestion, preprocessing, model training, deployment, and inference.

Data ingestion - Bird sound recordings are uploaded to AWS S3 from manual uploads.

Preprocessing - AWS Lambda functions are triggered to apply preprocessing operations including noise reduction, normalization, and segmentation to enhance audio quality.

Preprocessing orchestration - AWS Step Functions manage the coordination of preprocessing tasks, automate quality control checks, and ensure robust error handling.

Model training - AWS SageMaker trains a fine-tuned Wav2Vec model using labeled bird sound data.

Model deployment - The trained model is deployed as an SageMaker Endpoint for real-time classification.

Inference & Prediction - New recordings are classified, and predictions are stored in AWS DynamoDB. AWS CloudWatch logs performance metrics, and DynamoDB stores classification results for retrieval.

5. Data pipeline architecture

The architecture of this system comprises six sequential stages integrated through AWS serverless services. It is designed to automate the entire workflow—from data ingestion to real-time deployment of a Wav2Vec model using SageMaker Endpoint. This architecture was evaluated using the AWS Well-Architected Framework: Data Analytics Lens, which demonstrated that the system aligns with all five pillars: Operational Excellence, Security, Reliability, Performance Efficiency, and Cost Optimization.

Each component contributes to system robustness—automated orchestration and monitoring via AWS Step Functions and CloudWatch; secure access and role-based permissions across Lambda, S3,

and SageMaker; fault-tolerant execution and error recovery mechanisms; efficient use of serverless compute for dynamic workloads; and minimized cost by paying only for what is used.

These factors collectively confirm that the system is well-architected for scalable, secure, and real-time bird sound classification in cloud-based environments. It provides a practical, extensible model for bioacoustics research and biodiversity monitoring using serverless cloud solutions.

6.Result and Discussion

The classification performance was assessed using standard evaluation metrics: accuracy, precision, recall, and F1-score. The model demonstrated varying degrees of effectiveness across the three bird species. For the White-breasted Waterhen, the model achieved a precision of 0.61, recall of 0.88, and an F1-score of 0.72. The Black-browed Reed Warbler showed the highest performance with a precision of 1.00, recall of 0.64, and an F1-score of 0.78. In contrast, the Stork-billed Kingfisher presented the most challenging classification results, with a precision of 0.50, recall of 0.31, and an F1-score of 0.38.

These outcomes suggest that the model was more adept at recognizing the vocalizations of the Waterhen and Reed Warbler, likely due to their clearer and more consistent acoustic signatures. On the other hand, the lower performance for the Kingfisher may result from more diverse vocalization patterns or fewer high-quality training samples. The overall classification metrics, including a micro-average and weighted F1-score of 0.65, indicate a balanced yet moderate level of classification accuracy. Additional insight was provided by confusion matrix visualizations, which highlighted the areas where the model confused one species with another.

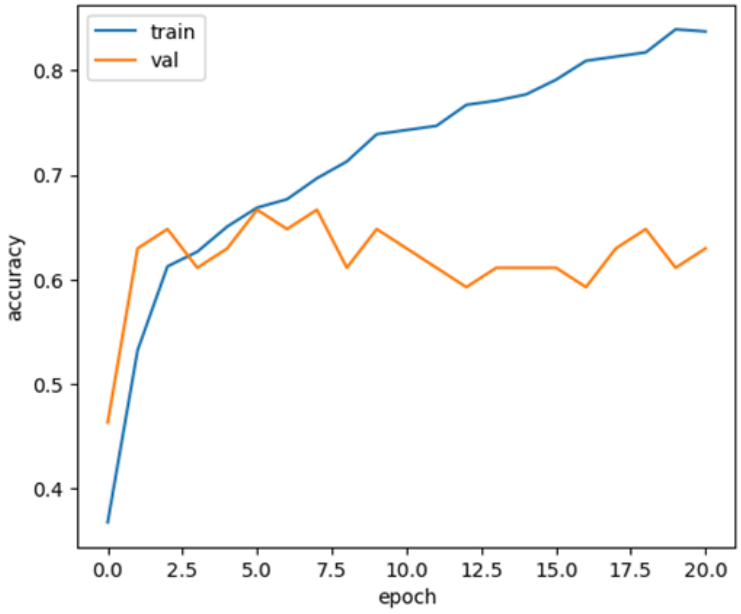

Figure 6 Training and validation accuracy

Figure 6 presents the training and validation accuracy of the Wav2Vec model over 20 training epochs. The training accuracy shows a steady upward trend, indicating that the model is successfully learning from the training data. In contrast, the validation accuracy initially increases but plateaus after approximately the 5th epoch and fluctuates thereafter. This pattern suggests that while the model continues to improve on the training set, its performance on unseen validation data does not

significantly improve beyond the early epochs. This divergence between training and validation accuracy is indicative of potential overfitting, where the model fits too closely to the training data and fails to generalize well to new inputs. Further strategies such as regularization, data augmentation, or tuning early stopping criteria could be employed to address this gap and enhance model generalization.

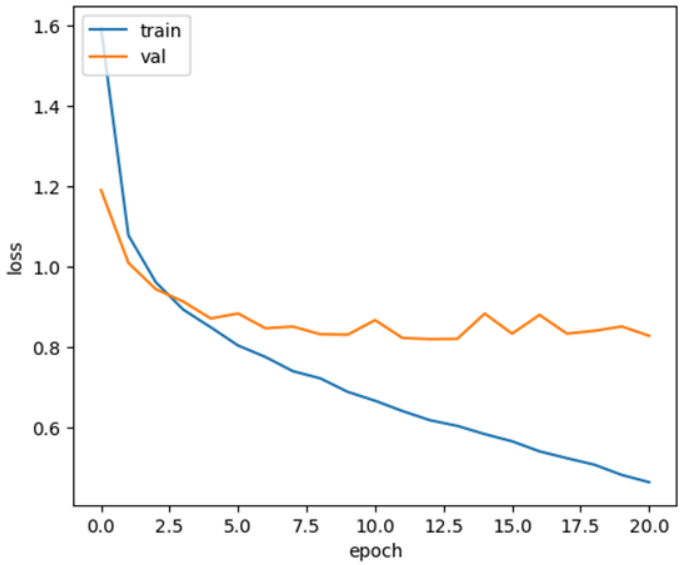

Figure 7 Training and validation loss

Figure 7 displays the training and validation loss curves across 20 epochs. The training loss steadily decreases throughout the training process, indicating that the model is consistently optimizing and fitting the training data. However, the validation loss decreases initially but stabilizes early and begins to fluctuate slightly after the 5th epoch. This divergence between training and validation loss suggests that the model starts to overfit the training data after a certain point, learning patterns that do not generalize well to unseen data. The increasing gap between the two loss curves reinforces the presence of overfitting. To mitigate this, techniques such as dropout, early stopping, or increased data augmentation could be considered in future iterations to improve model robustness and generalization.

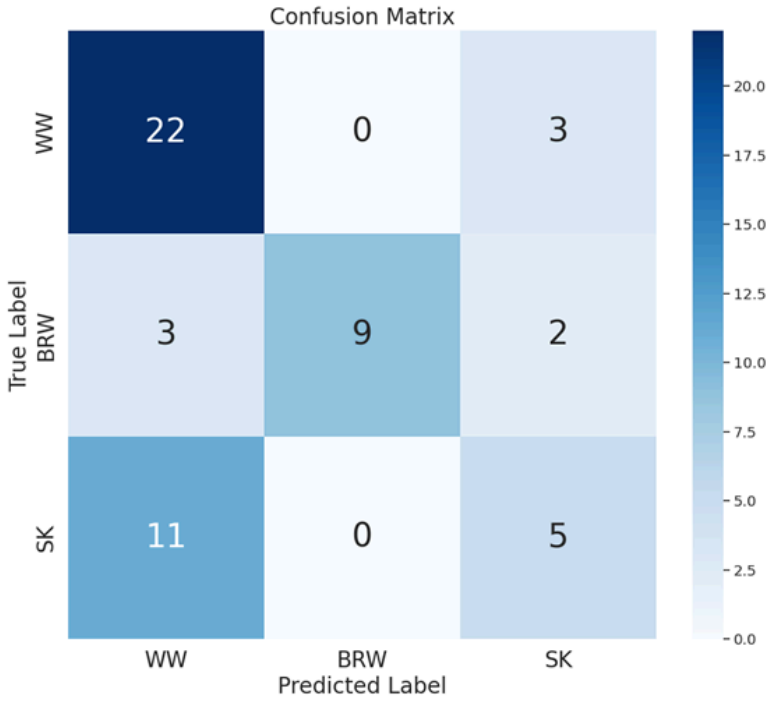

Figure 8 Confusion matrix

Figure 8 illustrates the confusion matrix for the bird sound classification model, showing how accurately the model predicted each class. The matrix compares the true labels (rows) with the predicted labels (columns) across three bird species. The first class (row 0) corresponds to the White-breasted Waterhen (WW), with 22 correct predictions, 3 misclassified as class 2, and none as class 1, indicating strong model performance on this class. The second class (row 1), representing the Black-browed Reed Warbler (BRW), had 9 correct predictions, but also 3 misclassified as class 0 and 2 as class 2, showing moderate confusion with other classes. The third class (row 2), the Stork-billed Kingfisher, (SK), was more problematic: only 5 out of 16 were correctly classified, while 11 were misclassified as class 0. This suggests that the model had difficulty distinguishing this species, potentially due to overlapping acoustic features or insufficient training samples.

Overall, the confusion matrix confirms the earlier quantitative findings—strong classification performance for the Waterhen, moderate for the Reed Warbler, and significant misclassification for the Kingfisher. These results highlight the need for targeted improvements, such as additional training data or model tuning for underperforming classes.

Table 4 Performance of the proposed using Wav2Vec

Type	precision	recall	f1-score	support
White-breasted Waterhen (WW)	0.61	0.88	0.72	25
Black-browed Reed Warbler (BRW)	1.00	0.64	0.78	14

Stork-billed Kingfisher (SK)	0.50	0.31	0.38	16
Micro average	0.65	0.65	0.65	55
Macro average	0.70	0.61	0.63	55
Weighted average	0.68	0.65	0.64	55
Sample verage	0.65	0.65	0.65	55

The performance results reveal variations in the F1-scores for the three bird species, as shown in Table 4: WW (0.72), BRW (0.78), and SK (0.36). These differences suggest species-specific challenges in sound classification. The relatively high F1-scores for the WW and BRW indicate that the model effectively identifies their vocalizations, likely due to the distinctiveness and consistency of their acoustic features. The results are further detailed in Figure 8, which presents the confusion matrix.

Conclusion

This study presents the successful development and deployment of a serverless data pipeline for bird sound classification using deep learning in a cloud-based environment. By leveraging a range of AWS services—namely Amazon S3 for data storage, AWS Lambda for preprocessing, Step Functions for orchestration, and SageMaker for model training and deployment—the proposed system implemented a fully automated pipeline. The integration of the Wav2Vec model enabled efficient processing of raw waveform audio without the need for extensive manual feature engineering, thereby streamlining the workflow.

The classification results showed that the model achieved satisfactory performance for two out of the three target bird species, indicating its applicability for biodiversity monitoring and ecological research. While this study did not capture quantitative cloud performance metrics such as processing time or cost, the functional deployment confirmed that the serverless architecture performed reliably across all pipeline stages. These findings underscore the practical viability of using serverless cloud environments for deploying machine learning models, particularly in research settings with limited access to high-performance computing infrastructure.

The successful operation of the pipeline on AWS demonstrates that serverless cloud solutions are not only technically feasible but also effective for scalable, event-driven applications. The end-to-end system was able to ingest, process, classify, and store bird sound data without manual server management, offering a replicable and extensible model for similar environmental monitoring tasks.

Future work

Future research may focus on several key areas to enhance and extend the capabilities of the current system:

- **Dataset Expansion and Quality Improvement**: Incorporating a larger and more diverse dataset, covering additional bird species and geographic regions, will help generalize the model and improve robustness. Special attention should be given to improving the quality of audio samples and ensuring balanced class representation.

- **Model Robustness and Optimization**: Investigating alternative model architectures that are more resilient to noise and capable of handling imbalanced or sparse data can improve classification performance. Techniques such as data augmentation, transfer learning, and ensemble methods may offer further improvements.

- **Real-Time IoT Integration**: Developing a smart sensor network using Internet of Things (IoT) technologies could enable real-time and continuous bird sound monitoring. This would allow for more dynamic data collection and support field-based ecological studies in remote areas.

- **User Accessibility and Interface Design**: Building a user-friendly graphical interface or mobile application would increase accessibility for conservationists, researchers, and citizen scientists. Such tools could facilitate real-time data upload, result visualization, and broader public engagement.

By addressing these directions, future iterations of this work can evolve into a more comprehensive, adaptive, and impactful system for bioacoustics research and environmental conservation.

Github link:

https://github.com/thitiwan13/cs653-bird-sound-classification

REFERENCES

1. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477

2. Bandara, M., Jayasundara, R., Ariyarathne, I., Meedeniya, D., & Perera, C. (2023). Forest Sound Classification Dataset: FSC22. *Sensors, 23*(4), 2032. https://doi.org/10.3390/s23042032

3. Dong, X., Towsey, M., Zhang, J., Banks, J., & Roe, P. (2013). A novel representation of bioacoustic events for content-based search in field audio data. *IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/ICASSP.2013.6691473

4. Feng, Z., Markov, K., Saito, J., & Matsui, T. (2024). Neural Cough Counter: A Novel Deep Learning Approach for Cough Detection and Monitoring. IEEE Access, 12, 118816–118829. https://doi.org/10.1109/ACCESS.2024.3449370

5. Höchst, J., Bellafkir, H., Lampe, P., Vogelbacher, M., Mühling, M., Schneider, D., Lindner, K., Rösner, S., Schabo, D. G., Farwig, N., & Freisleben, B. (2022). Bird@Edge: Bird species recognition at the edge. In *Machine Learning for Biodiversity Research* (pp. 101-118). Springer. https://doi.org/10.1007/978-3-031-17436-0_6

6. Mbata, A., Sripada, Y., & Zhong, M. (2024). A survey of pipeline tools for data engineering. *arXiv*. https://arxiv.org/abs/2406.08335

7. Mou, A., & Milanova, M. (2024). Performance analysis of deep learning model-compression techniques for audio classification on edge devices. *Sci, 6*(2), 21. https://doi.org/10.3390/sci6020021

8. Noumida, A., & Rajan, R. (2022). Multi-label bird species classification from audio recordings using attention framework. *Applied Acoustics, 197*, 108901. https://doi.org/10.1016/j.apacoust.2022.108901

9. Shojaee Rad, Z., & Ghobaei-Arani, M. (2024). Data pipeline approaches in serverless computing: A taxonomy, review, and research trends. *Journal of Big Data, 11*, 82. https://doi.org/10.1186/s40537-024-00939-0

10. Sigtia, S., Stark, A. M., Krstulović, S., & Plumbley, M. D. (2016). Automatic environmental sound recognition: Performance versus computational cost. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(11), 2096-2107. https://doi.org/10.1109/TASLP.2016.2592698

11. Swaminathan, B., Jagadeesh, M., & Subramaniyaswamy, V. (2024). Multi-label classification for acoustic bird species detection using transfer learning approach. *Ecological Informatics, 80*, 102471. https://doi.org/10.1016/j.ecoinf.2024.102471

12. Tan, K. I., Yean, S., & Lee, B. S. (2023). Attention-based sound classification pipeline with sound spectrum. *IEEE Sensors Applications Symposium (SAS)*. https://doi.org/10.1109/SAS58821.2023.10254193

13. Wimmer, J., Towsey, M., Planitz, B., Roe, P., & Williamson, I. (2010). Scaling acoustic data analysis through collaboration and automation. *Proceedings of the Sixth IEEE International Conference on e–Science*, 308–315. https://doi.org/10.1109/eScience.2010.17