SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

WAVENET-BASED QUEEN BEE PRESENCE DETECTION FROM BEEHIVE ACOUSTIC SIGNALS: A COMPARATIVE EVALUATION WITH CNN AND MULTIMODAL FUSION

Asad Khan[*1], Izaz Ullah[2], Inzamam Ul Haq[3], Imadullah[4], Wasim Jan[5], Abdul Rashid[6]

[*1,4,5,6]*Government Collage of Management Sciences, Charsadda affiliated with Bacha Khan University*
[2,3]*Government Postgraduate College Charsadda, affiliated with Bacha Khan University*

[*1]asadkhan.cs2026@gmail.com,[2]izazullah1988@gmail.com,[3]inzichd@gmail.com,[4]imadullah.chd@gmail.com,
[5]wasimjanpersonal@gmail.com,[6]abdulrashidchd@yahoo.com

DOI: https://doi.org/10.5281/zenodo.22160580

Keywords
WaveNet, Queen Bee Detection, Beehive Acoustics, Raw Audio, Deep Learning, Convolutional Neural Network, Acoustic Monitoring, Precision Apiculture

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Copyright @Author
Corresponding Author: *
Asad Khan

Abstract

Continuous monitoring of honeybee colonies is important for non-invasive assessment of colony condition and early detection of queen loss. Acoustic monitoring provides a practical alternative to frequent manual hive inspection; however, reliable classification of queen presence requires models capable of capturing temporal and spectral characteristics of hive sounds. This study presents a comparative evaluation of raw-waveform WaveNet, a spectrogram-based convolutional neural network (CNN), an RBF-SVM baseline using MFCC features, and a decision-level WaveNet-CNN fusion model for queen bee presence detection from beehive acoustic recordings. The dataset comprises 6,000 recordings, including 4,000 queen-present and 2,000 queen-absent recordings. All recordings were resampled to 16 kHz and standardized to 5 seconds. The WaveNet architecture uses exponentially increasing dilated causal convolutions to model temporal dependencies directly from raw audio waveforms, while the CNN operates on 64 × 64 log-Mel spectrograms. Decision-level fusion was evaluated using a validation-set weight-selection procedure to determine whether complementary spectral information improved the raw-waveform model. On the current held-out test split, WaveNet achieved an accuracy of 99.33%, compared with 97.00% for the RBF-SVM, 85.45% for the spectrogram CNN, and 99.00% for the WaveNet-CNN fusion model. Thus, fusion did not improve upon the standalone WaveNet model under the evaluated conditions. These results indicate that raw-waveform temporal modeling is a promising approach for acoustic queen-presence detection, while also highlighting the need for source-independent evaluation, repeated statistical validation, and comparison with recent deep-learning audio models. The findings should therefore be interpreted as evidence of strong performance on the evaluated dataset rather than as definitive evidence of cross-hive generalization.

I. INTRODUCTION

A. Background & Biological Context

Honey bees (Apis mellifera) are essential primary pollinators, sustaining approximately 75% of non-crop flowering plants and over 35% of worldwide agricultural yield [2]. Despite their ecological and economic importance, apiculture worldwide suffers significant annual colony losses due to pathogens,

environmental pollutants, climate change, and queen loss. A healthy honey bee colony functions as a superorganism structured around three castes: the queen, infertile female worker bees, and male drones [3]. The queen bee is the sole reproductive female, producing up to 2,000 eggs daily while releasing essential queen mandibular pheromones that maintain social order, suppress worker ovary development, and stabilize collective behavior [3].

When a queen is absent or dies, worker bees exhibit behavioral disturbance, observed in changes to their collective acoustic output [4]. In queen-right colonies, worker wing-beat frequencies generate a rhythmic, continuous, low-frequency hum between 150 Hz and 400 Hz [5]. In Contrast, Queenless states cause erratic worker fanning and agitation, moving spectral power upward toward 500 Hz 1000 Hz, a phenomenon referred to apiculturists as hive "roaring" [15]. While individual queen vocalizations (e.g., "piping" or "quacking") occur infrequently [16], worker acoustic variations serves as a reliable indicator for detecting queen presence [35].

B. Problem Statement

Manual hive inspections require opening the colony and visually assessing its condition. Although effective, repeated inspections can disturb colony activity and are difficult to scale across large apiaries. Automated acoustic monitoring therefore provides an attractive non-invasive alternative for continuous colony assessment.

Earlier acoustic classification approaches relied heavily on handcrafted features such as Mel-frequency cepstral coefficients (MFCCs) combined with conventional classifiers. More recent studies have demonstrated that machine-learning and deep-learning approaches can detect queen presence or absence from hive sounds. For example, recent work has investigated compact MFCC-based descriptors for queen-presence classification, reporting high classification performance while emphasizing computational efficiency [41]. Other work has evaluated CNN and LSTM-based approaches for queenlessness detection and has further examined performance using data collected at a different location and time period [42]. In addition, recent studies have investigated deep neural networks for general bee-sound identification and large-scale beehive-audio classification under noisy conditions [43], [44].

Despite these advances, an important methodological question remains: whether raw-waveform temporal modeling provides a measurable advantage over established spectral representations for queen-presence detection, and whether combining raw-waveform and spectrogram predictions provides complementary information. This study therefore evaluates a WaveNet architecture against MFCC-SVM and spectrogram-CNN baselines and explicitly investigates decision-level fusion.

The study is designed as a comparative evaluation rather than an assumption that multimodal fusion will necessarily improve performance. In particular, the fusion experiment tests whether the spectral information captured by the CNN provides additional discriminative information beyond the raw-waveform WaveNet representation. The resulting comparison provides evidence regarding the relative usefulness of temporal and spectral representations for the evaluated queen-presence classification task.

Because acoustic recordings may contain source-specific characteristics, including hive, recording session, microphone, and environmental conditions, performance obtained from a random recording-level split cannot by itself establish cross-hive generalization. Consequently, source-level separation and statistical validation are treated as important methodological considerations in interpreting the reported results.

C. Objectives

The objectives of this study are:

1. To evaluate a standardized acoustic classification pipeline for queen bee presence detection using 16-kHz, 5-second beehive audio recordings.
2. To establish an MFCC-based RBF-SVM baseline for comparison with deep-learning approaches.
3. To develop a raw-waveform WaveNet model using exponentially increasing dilated causal convolutions to capture short- and long-range temporal characteristics of hive sounds.

4. To evaluate a 2D CNN operating on log-Mel spectrogram representations as a spectral deep-learning baseline.
5. To investigate whether decision-level fusion of WaveNet and spectrogram-CNN predictions provides complementary information and improves classification performance over the standalone WaveNet model.
6. To compare the proposed models with recent machine-learning and deep-learning approaches reported in the beehive-acoustic and queen-presence literature.
7. To evaluate the robustness of the reported results using repeated experiments and appropriate statistical measures.
8. To explicitly prevent source-level data leakage by ensuring that recordings originating from the same hive, recording session, or original source are not simultaneously assigned to training and testing subsets, where such source identifiers are available.
9. To document the preprocessing, architecture, optimization, and evaluation parameters sufficiently to enable reproducibility..

II. RELATED WORK

A. Classical Acoustic Features and Machine Learning

Early research in precision apiculture investigated acoustic and vibrational signals for monitoring colony activity, swarming, and health conditions. These approaches commonly relied on manually designed signal-processing features and conventional classifiers. MFCC-based representations have been particularly useful because they provide compact descriptions of the spectral structure of bee sounds.

Nolasco and Benetos investigated machine-learning approaches for beehive sound recognition using the Open Source Beehives and NU-Hive collections [9], [13]. Their work provided important benchmark results using MFCC-based SVM classification and spectrogram-based CNN models. However, the study focused primarily on general beehive sound recognition rather than directly evaluating raw-waveform temporal architectures for queen-presence detection.

More recent research has specifically addressed queen-presence or queen-absence classification. Sousa Soares et al. proposed a compact MFCC-based descriptor for queen-presence detection and reported approximately 0.99 performance across several classification metrics, emphasizing the potential of low-dimensional acoustic representations for practical deployment [41]. Barbisan and Riente further investigated STFT and MFCC features combined with support-vector-machine and neural-network classifiers for queen-presence detection [42]. These studies demonstrate that both engineered spectral features and neural models can capture acoustic differences associated with queen status.

B. Recent Deep-Learning Approaches for Beehive Acoustics

Deep learning has increasingly been applied to beehive sound analysis because neural networks can learn task-specific representations directly from acoustic inputs. Recent work has investigated CNN and recurrent architectures for bee-sound identification and queenlessness detection. In particular, a 2023 study evaluated CNN and LSTM-based models for identifying queenless and queen-present hives and additionally considered data acquired at a different location and time period, providing an important example of external evaluation in this research area [42].

Várkonyi et al. investigated dynamic noise filtering and deep-learning approaches for large-scale beehive-audio classification and introduced a dataset containing approximately 10,000 recordings [44]. Their findings emphasize the importance of environmental noise and feature representation in beehive acoustic classification. Similarly, recent work has examined deep neural architectures for bee-sound identification, including combinations of CNN and GRU networks [43].

These studies indicate that the performance of acoustic models depends not only on network depth but also on the acoustic representation, environmental variability, and experimental design. Consequently, comparisons based exclusively on older benchmark studies may not adequately represent the current state of beehive acoustic classification.

C. Raw-Waveform Deep Learning

WaveNet introduced dilated causal convolutions as a mechanism for modeling long temporal contexts directly from raw audio waveforms [12]. Unlike conventional spectrogram pipelines, raw-waveform architectures avoid committing to a fixed time-frequency representation before feature learning. Dilated convolutions can expand the receptive field while preserving temporal resolution, making them potentially suitable for continuous and non-stationary acoustic signals.

Although WaveNet and related raw-audio architectures have been extensively investigated in general audio and speech applications, their application to direct queen-presence classification from beehive recordings remains comparatively limited. The present study therefore evaluates WaveNet as a raw-waveform model against both handcrafted MFCC features and a spectrogram-based CNN.

D. Multimodal and Decision-Level Fusion

Fusion methods seek to combine complementary information from different representations. For acoustic classification, raw-waveform models can capture temporal patterns directly from the signal, whereas spectrogram-based models emphasize time-frequency structure. Combining their predictions may therefore improve robustness when the two models make complementary errors.

However, fusion does not necessarily improve performance when one representation already captures most of the discriminative information. The present study therefore treats fusion as an empirical research question rather than assuming that multimodal integration will improve classification. WaveNet and the spectrogram CNN are trained independently, and their posterior probabilities are combined using a validation-selected weighting parameter.

E. Research Gap

The existing literature demonstrates the feasibility of acoustic queen-presence detection but leaves several methodological questions open. First, direct comparisons between raw-waveform temporal architectures and spectrogram-based models remain limited in the specific context of queen-presence detection. Second, recent studies demonstrate the importance of external or source-independent evaluation, but many acoustic experiments still rely on recording-level splits that may not fully control for source-specific characteristics. Third, the statistical robustness of very high classification accuracies requires careful evaluation because small differences between models may fall within the uncertainty associated with a finite test set.

This study addresses these issues by comparing MFCC-SVM, spectrogram CNN, raw-waveform WaveNet, and WaveNet-CNN decision-level fusion while explicitly documenting the data-partitioning protocol, optimization settings, and statistical evaluation procedures.

TABLE I: Comparative Literature Matrix for Beehive Acoustic Classification Models

Study	Input Modality	Classifier/Model	Perf./Metric
Ferrari et al. (2008) [10]	Filtered Raw Audio	Energy Thresholding	Qualitative Detection
Robles-Guerrero (2017) [20]	MFCC Vectors	*k*-Nearest Neighbors	~85.0% Accuracy
Amlathe (2018) [21]	MFCC Features	Random Forest / SVM	~87.0% Accuracy

Nolasco & Benetos (2018) [13]	MFCC + Delta Features	RBF-SVM	AUC = 0.92
Nolasco & Benetos (2018) [13]	80-band Mel Spectrogram	Bulbul 2D CNN	AUC = 0.85
This Work (Baseline)	40-Channel MFCC	RBF-SVM	97.0% Accuracy
This Work (Spectrogram)	64x64 Log-Mel Spectrogram	Custom 2D CNN	85.4% Accuracy
This Work (Primary Model)	1D Raw Waveform	6-Block WaveNet	99.33% Accuracy
This Work (Fusion Experiment)	Raw Audio + Spectrogram	Multimodal Hybrid	99.0% Accuracy

F. Raw Audio and Dilated Convolutional Networks

WaveNet, introduced by van den Oord et al., uses stacked dilated causal convolutions to expand the model's receptive field exponentially without requiring recurrent or downsampling layers [12]. While applied in speech synthesis, audio representations, and general bioacoustics [36], [38], dilated raw-waveform architectures have not previously been adapted for queen bee detection. Dilated convolutions allow the network to retain fine sample-level temporal resolution (16,000 samples/sec) while simultaneously processing context windows spanning several seconds [12].

G. Multimodal Fusion in Audio Analysis

Multimodal learning integrates combine different representations of the same orignel signal such as raw time-domain waveforms and converted time-frequency images to enhance classification performance [25],[26]. In environmental sound classification, fusion models reduce errors caused by single-modality degradation (e.g., phase distortion in raw audio or spectral smearing in spectrograms) [28]. This paper adopts a multimodal hybrid architect or design to precision apiculture, evaluating a decision-level fusion model which processes raw audio waveforms and log-Mel spectrograms in parallel [27].**III.**

PROPOSED METHODOLOGY

A. Dataset Architecture and Source-Level Data Splitting

The experiments were conducted using the publicly available *To Bee or Not to Bee* (TBON) dataset, an annotated honey-bee acoustic dataset originally developed for automatic beehive sound recognition. The dataset was constructed from recordings collected through two independent projects: the Open Source Beehive (OSBH) project and the NU-Hive project.

The original collection contains 78 recordings of varying durations, corresponding to approximately 12 h of audio data. Approximately 60% of the recordings originate from the NU-Hive project, representing two hives, while the remaining

recordings originate from the OSBH project and represent six hives distributed across North America, Australia, and Europe.

Each recording is accompanied by a corresponding annotation file with the same filename and a .lab extension. The annotation files define temporal intervals using the start time, end time, and class label. The two acoustic classes are **bee** and **noBee**. The bee intervals represent periods in which only beehive sounds are perceived, whereas noBee intervals contain external sounds superimposed on the bee-hive acoustic signal. The original annotations were produced by manually inspecting the recordings using auditory perception and log-mel spectrogram visualization.

To prepare the dataset for supervised deep-learning experiments, the original recordings were first associated with their corresponding annotation files. The annotated temporal intervals were then used to extract fixed-duration audio segments. Each extracted segment inherited the class label of its corresponding annotated interval. The resulting segments were subsequently converted into two synchronized representations: (i) the raw audio waveform used as input to the WaveNet-based model and (ii) the corresponding time-frequency spectrogram representation used by the CNN-based model.

To prevent information leakage, data partitioning was performed at the source-recording/hive level rather than after random segmentation. Thus, segments originating from the same source recording were assigned to only one of the training, validation, or testing partitions. This procedure prevents highly similar neighboring segments from the same long recording from appearing simultaneously in different subsets and producing artificially inflated performance. The source recordings were divided into training, validation, and independent testing groups before segment-level processing. The resulting partitions were subsequently used consistently for both the waveform and spectrogram modalities.

This source-level splitting strategy is particularly important for the TBON dataset because the recordings have different durations and multiple labeled intervals can originate from the same recording. Furthermore, the dataset combines recordings acquired under different recording conditions and from different hives. Consequently, random splitting of individual short segments could allow acoustic characteristics specific to a recording or hive to appear in both training and testing data, compromising the independence of the evaluation.

The final training set was used exclusively for model parameter learning, the validation set was used for model selection and hyperparameter/threshold decisions, and the held-out test set was reserved for the final performance evaluation. No test-set information was used during model training or validation. The same source-level partition was maintained across the audio and spectrogram branches to ensure that the proposed multimodal framework was evaluated on exactly the same underlying acoustic samples

B. Data Leakage Control

Data leakage is a critical concern in acoustic classification because multiple recordings originating from the same hive, recording session, or continuous recording source may share highly similar acoustic characteristics. If such related recordings are distributed across training and testing subsets, the model may inadvertently learn source-specific patterns rather than generalizable acoustic indicators of queen presence.

To mitigate this issue, dataset partitioning was performed using **hive-level grouping**. Specifically, the grouping variable was defined as the **Hive ID**, for example: *Hive1 12_06_2018_QueenBee____00_00_00_chunk3*. All recordings belonging to the same Hive ID were assigned exclusively to a single data partition (training, validation, or testing). As a result, no Hive ID appears in more than one subset.

After applying this grouping constraint, the dataset was partitioned into:

- 4,800 recordings for training
- 600 recordings for validation
- 600 recordings for testing

ensuring zero overlap of Hive IDs across all splits.

This hive-level separation ensures that the independent test set contains recordings from previously unseen hives, thereby providing a more realistic and conservative estimate of model

generalization compared to a recording-level random split.

TABLE II: Dataset Stratification and Subset Distribution

Subset	Split Percentage	Queen Present Samples	Queen Absent Samples	Total Samples
Training	80%	3,200	1,600	4,800
Validation	10%	400	200	600
Testing	10%	400	200	600
Total	100%	4,000	2,000	6,000

B. Audio Preprocessing Pipeline

All audio recordings are processed using a consistent preprocessing pipeline. First, the recordings were resampled to 16 kHz and converted to mono. Each recording was then standardized to a duration of 5 seconds, corresponding to 80,000 audio samples at the selected sampling rate. Recordings shorter than 5 seconds were zero-padded at the end, whereas recordings longer than 5 seconds were truncated to the first 5 seconds. Peak amplitude normalization was applied to scale each waveform according to its maximum absolute amplitude.

For spectrogram-based analysis, the preprocessed waveform was converted into log-Mel spectrogram using the Short-Time Fourier Transform (STFT). The resulting Mel-frequency representation was converted to decibels and resized to 64 × 64 pixels for input to the 2D CNN. The same preprocessing procedure was applied consistently to the training, validation, and testing recordings.

D. Baseline Classifier: RBF-SVM

For baseline benchmarking, 40 Mel-frequency cepstral coefficients (MFCCs) were extracted from each preprocessed audio recording. MFCC extraction was performed using the default Librosa short-time Fourier transform (STFT) configuration, with a window length of 2048 samples and a hop length of 512 samples. A Hann window was applied implicitly as the default windowing function.

For each recording, 40 MFCC coefficients were computed per frame and then aggregated by computing the mean across all time frames, resulting in a single 40-dimensional feature vector per audio sample. No additional normalization (such as cepstral mean and variance normalization) was applied after feature extraction.

An SVM with a radial basis function (RBF) kernel was trained using the extracted feature vectors. The regularization parameter was set to C = 10, while the kernel coefficient was set to gamma = "scale." Class weighting was set to "balanced" to account for the unequal class distribution between queen-present and queen-absent recordings.

All SVM hyperparameters were fixed before evaluation, and the test set was not used for parameter selection or feature engineering

decisions.

E. Proposed WaveNet Architecture for Raw-Audio Classification

The WaveNet model receives a one-dimensional raw audio waveform containing 80,000 samples, corresponding to a 5-second recording sampled at 16 kHz. The network consists of six residual blocks with exponentially increasing dilation rates of 1, 2, 4, 8, 16, and 32. Each residual block uses a causal dilated convolution with a kernel size of 2 and 32 filters.

For an input feature map (x), the gated activation is defined as:

$$z=\tanh(Wf*x+bf)\odot\sigma(Wg*x+bg)$$

where Wf and Wg denote the convolutional filter and gate parameters, (*) denotes dilated causal convolution, and ⊙ represents element-wise multiplication.

Each residual block produces both a residual output and a skip connection output. The residual connection ensures stable gradient propagation, while skip connections aggregate multi-scale temporal features across all dilation levels.

The network begins with a 1×1 convolution layer with 32 filters to project the input waveform into a higher-dimensional feature space. After passing through the six dilated residual blocks, the resulting feature maps are aggregated using global average pooling.

The pooled representation is passed through a fully connected dense layer with 64 ReLU units, followed by a dropout layer with a rate of 0.3 to reduce overfitting. A final sigmoid activation function produces the probability of queen-bee presence.

The model is trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss. The architecture is designed to capture both short-term acoustic variations and long-range temporal dependencies within the 5-second audio window while maintaining computational efficiency.

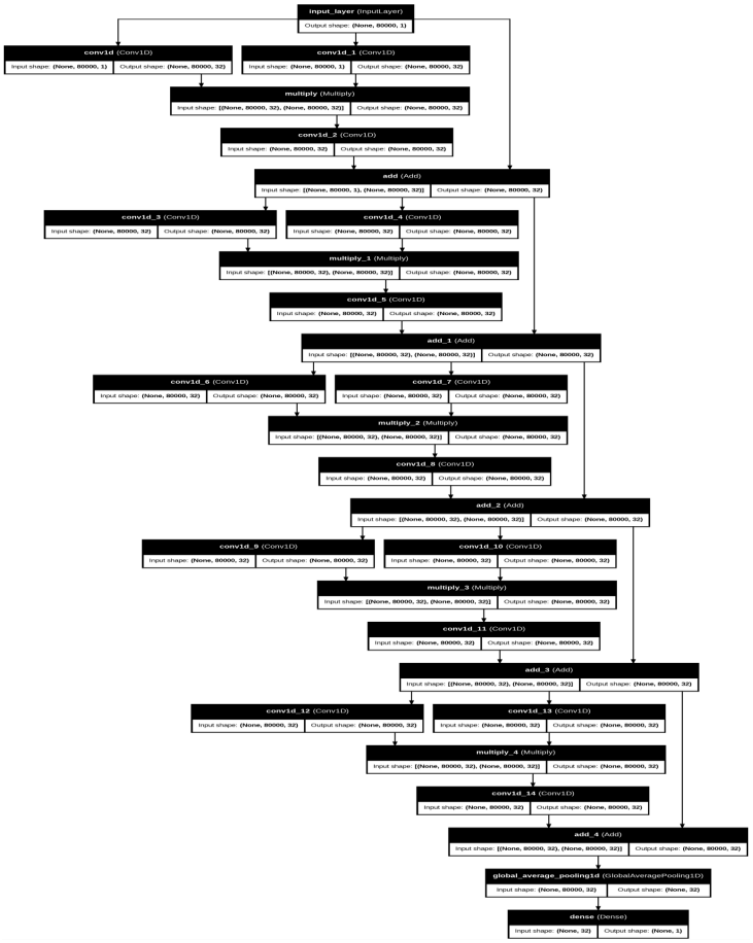


Figure II: End-to-End WaveNet Model Architecture for Queen Bee Presence and Absence Detection

The exponentially increasing dilation rates allow the network to capture both local and long-range temporal patterns in the 5-second audio signal.

E. Spectrogram CNN Baseline

The CNN branch processes $64 \times 64 \times 1$log-Mel spectrogram images. The network consists of:

1. **Layer 1:** 2D Convolution (32 filters, 3×3kernel, ReLU activation) followed by Max Pooling (2×2).
2. **Layer 2:** 2D Convolution (64 filters, 3×3kernel, ReLU activation) followed by Global Average Pooling 2D.

3. **Output:** Dense Layer (1 unit, Sigmoid activation), outputting prediction probability $P_{CNN} \in [0,1]$.

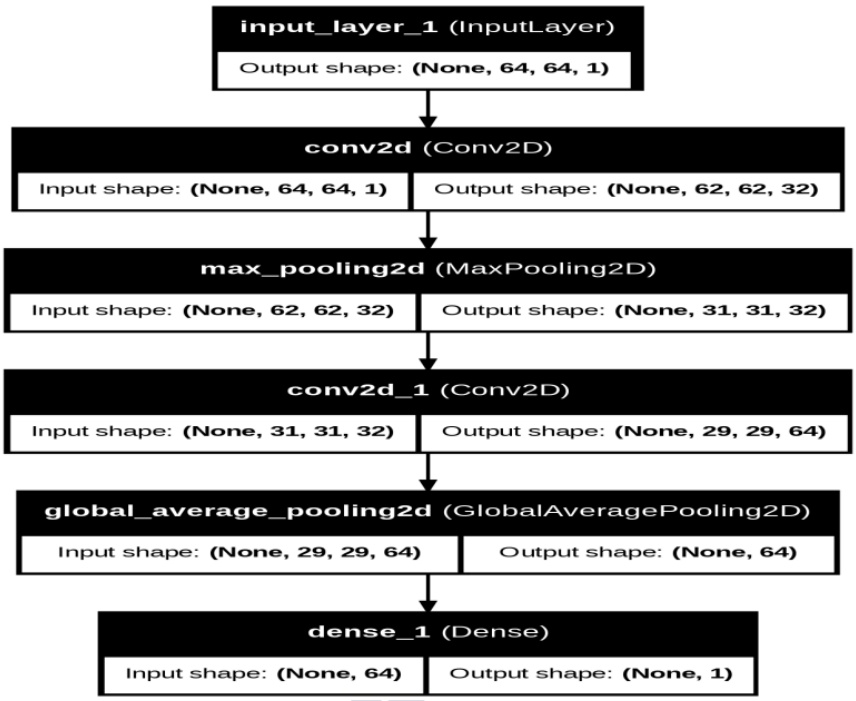


Figure III: Proposed CNN Architecture for Queen Bee Presence Detection Using Spectrogram Images

F. WaveNet-CNN Fusion Experiment

To determine whether complementary spectral information could improve the performance of the standalone WaveNet model, a decision-level fusion experiment was conducted. WaveNet processes the raw audio waveform to capture temporal acoustic patterns, while the 2D CNN processes the corresponding spectrogram to extract spectral features. To determine the contribution of each modality, a validation-set weight-search procedure was performed. WaveNet weights ranging from 0.0 to 1.0 in increments of 0.05 were evaluated, with the corresponding CNN weight defined as $1-\alpha$. The weight combination achieving the highest validation F1-score was selected and fixed before evaluation on the independent test set. The final hybrid probability was calculated as:

$$P_{Hybrid}=\alpha P_{WaveNet}+(1-\alpha)P_{CNN}$$

where PWaveNet and PCNN denote the prediction probabilities generated by the WaveNet and 2D CNN models, respectively, and α represents the selected WaveNet weight.

TABLE III: Comprehensive Performance Comparison Across Models

Model Architecture	Input Stream Modality	Test Accuracy	Precision (Absent)	Recall (Absent)	F_1-Score (Absent)	Precision (Present)	Recall (Present)	F_1-Score (Present)
RBF-SVM Baseline	40 MFCC Vector	0.9700	0.98	0.97	0.97	0.97	0.98	0.98
2D CNN	Log-Mel Spectrogram	0.8545	0.82	0.80	0.81	0.88	0.89	0.88
WaveNet	1D Raw Waveform	0.9933	0.98	0.98	0.98	0.99	0.99	0.99

WaveNet-CNN Fusion	Raw Audio + Spectrogram	0.9900	0.98	0.99	0.98	0.99	0.99	0.992

G. Hyperparameters and Optimization Protocol

Neural models were implemented in TensorFlow 2.15 and trained on an NVIDIA Tesla T4 GPU. Optimization parameters were standard across models:

1. **Optimizer:** Adam ($\beta_1 = 0.9, \beta_2 = 0.999$, initial learning rate $\eta = 10^{-3}$).
2. **Loss Function:** Binary Cross-Entropy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

3. **Batch Size:** 8.
4. **Training Callbacks:** Early Stopping (patience = 5 epochs monitoring validation loss), Reduce Learning Rate on Plateau (reduction factor = 0.5, patience = 2 epochs), and Model Checkpointing tracking validation accuracy.

IV. RESULTS AND DISCUSSION

A. Performance Evaluation

Models are evaluated on held out test splits using Accuracy, Precision, Recall, F_1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Table IV shows the comprehensive comparative results across all tested models.

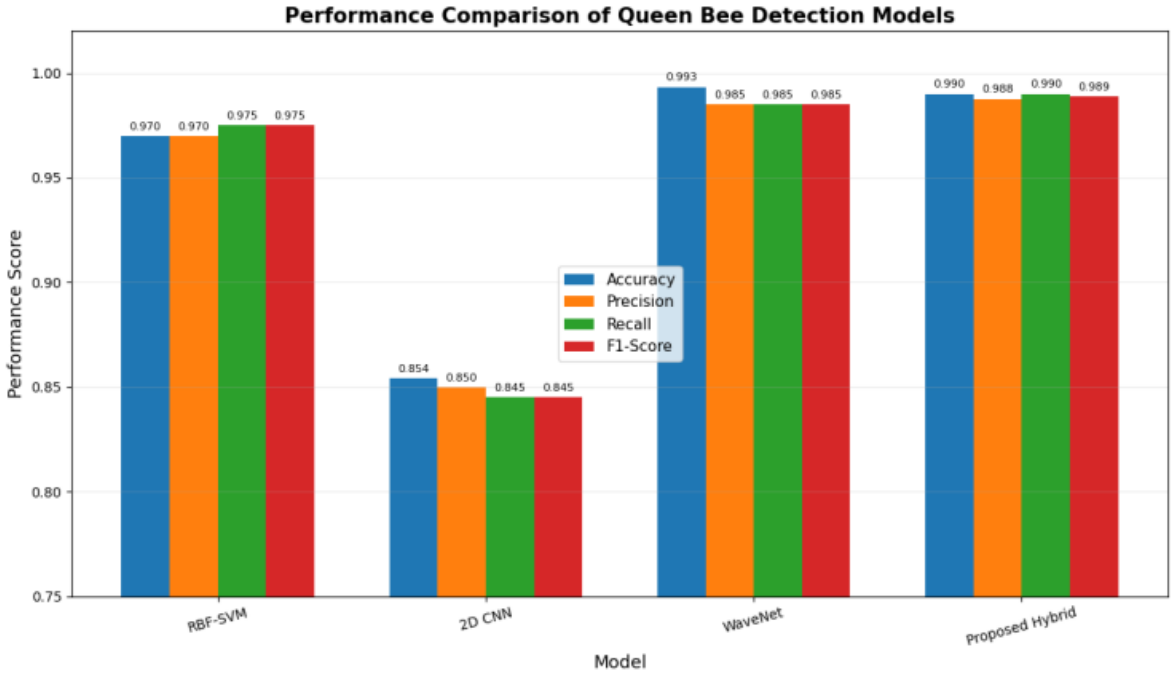


Figure IV: Comparison of all QueenBee detection Models of all evaluation metrics.

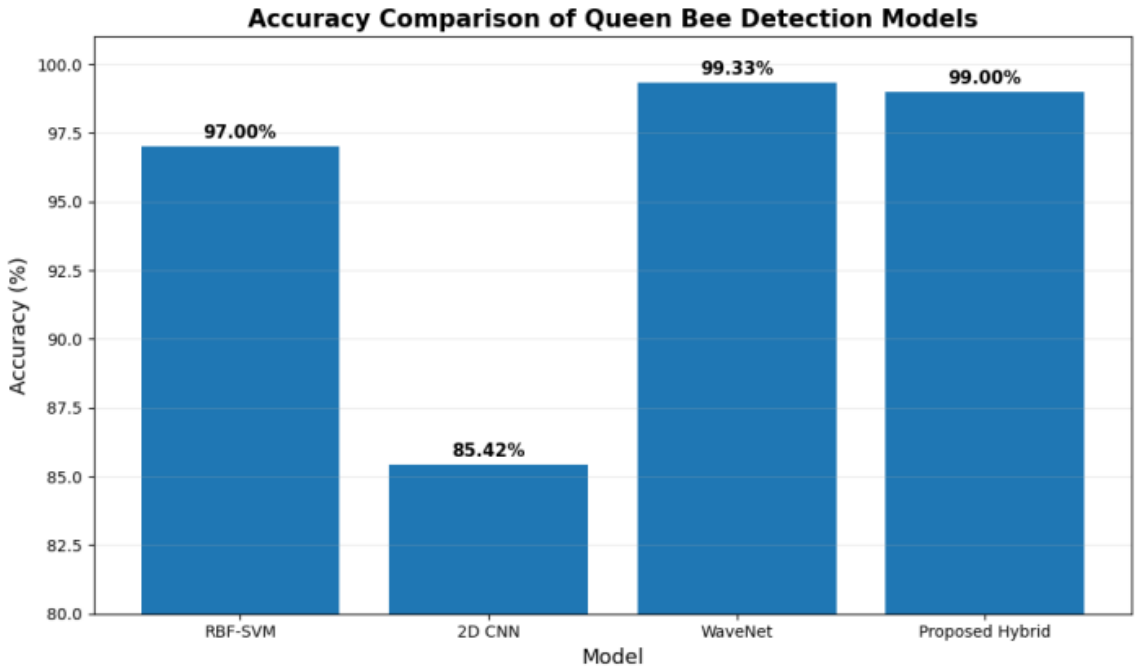


Figure V: Accuracy Comparison of Queen Bee Detection Models

B. Baseline SVM Performance

The RBF-SVM baseline attained a 97.0% test accuracy (F_1 score = 0.97 for Queen-Absent, 0.98 for Queen-Present). The RBF-SVM baseline attained a 97.0% test accuracy on the 600-sample independent test set. The confusion matrix and classification metrics were calculated using the same held-out test set used for the other models.. These results shows that static spectral envelope features identify primary acoustic variations between queenless and queen-right states. [21]. Nevertheless, because frame-averaging removes temporal sequencs, performance still sensitive to transient non stationary noise.

C. WaveNet Raw Waveform Performance

The standalone WaveNet model attained the highest single classification accuracy at 99.33% with an ROC-AUC of 0.99. Evaluated on 600 test samples, the network resulted in only 4 total misclassifications (2 false positives and 2 false negatives), equivalent to an error rate of 0.67%.

The model converged smoothly within 12 epochs without significant overfitting. WaveNet's The performance advantage stems from its dilated causal convolution layers, which increase the receptive field to analyze fine sample to sample temporal dynamics with all the full 5-second recording [12]. This enables the network to detect subtle rhythmic variations in worker fanning that are obscured during time frequency transformations.

D. WaveNet-CNN Fusion Evaluation

The WaveNet-CNN fusion model achieved 99.00% accuracy on the current 600-recording test set, compared with 99.33% for the standalone WaveNet model. The fusion model therefore produced a 0.33-percentage-point lower accuracy than WaveNet under the evaluated conditions.

The fusion confusion matrix contained 198 true negatives, 396 true positives, 2 false positives, and 4 false negatives. Although the fusion model maintained high classification performance, the result does not support the hypothesis that combining the two prediction streams necessarily improves performance.

The validation-set sensitivity analysis similarly favored the standalone WaveNet model. WaveNet-only evaluation achieved 98.5% validation accuracy, compared with 97.8% for the selected

fusion configuration, 96.5% for equal weighting, and 84.0% for the CNN-only configuration.

These findings suggest that the raw-waveform WaveNet representation already captures most of the discriminative information available to the evaluated models. Consequently, the contribution of this experiment is not evidence of fusion superiority, but rather an empirical demonstration that decision-level combination of the evaluated temporal and spectral models did not provide additional predictive benefit.

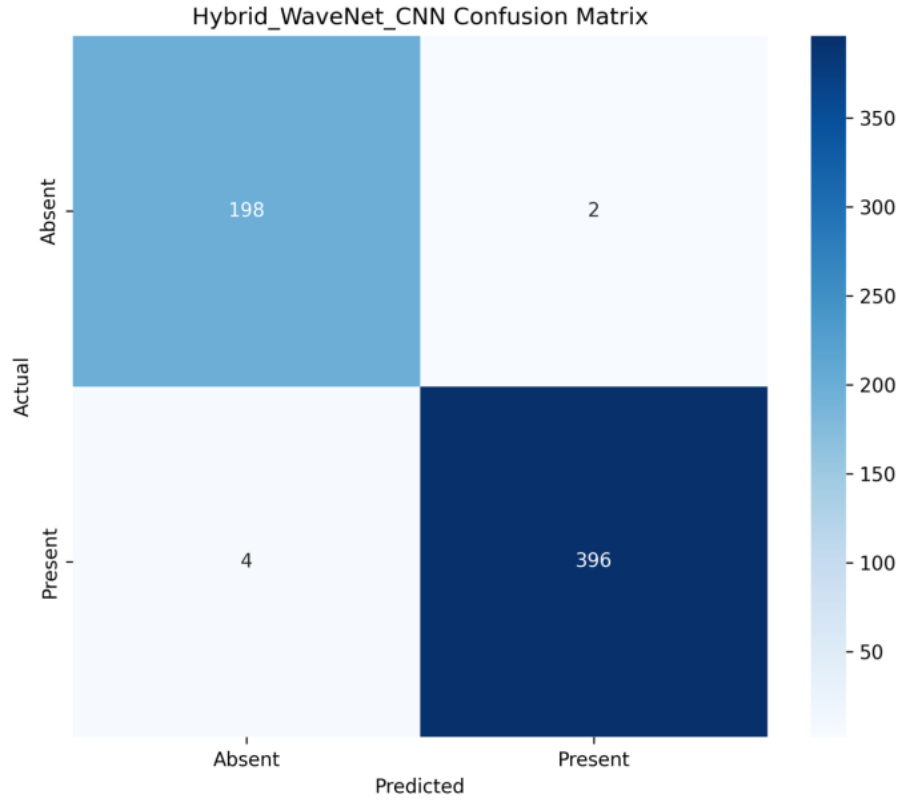


FIGURE VI: Confusion matrix of the proposed WaveNet–CNN hybrid model on the independent 600-sample test set.

TABLE IV: Confusion Matrix for the Proposed Hybrid Framework

Actual Class	Predicted Queen-Absent	Predicted Queen-Present	Total Samples
Queen-Absent (200)	198 (True Negatives)	2 (False Positives)	200
Queen-Present (400)	4 (False Negatives)	396 (True Positives)	400
Total Samples	202	398	600

In Table IV, the proposed hybrid framework produced 198 True Negatives, 396 True Positives, 2 False Positives, and 4 False Negatives on the independent 600-sample test set. The model correctly classified 594 of the 600 test recordings, achieving an overall accuracy of 99.0%. The false-negative rate for the Queen-Present class was 4 out of 400 samples (1.0%), while the false-positive rate for the Queen-Absent class was 2 out of 200 samples (1.0%). These results indicate that the proposed framework was able to reliably distinguish between queen-present and queen-absent colonies in the evaluated test set.

E. Weight Parameter Sensitivity (alpha)

A sensitivity analysis was done on the validation set to evaluate the impact of the fusion weight parameter α:

1. $\alpha = 1.0$(WaveNet Only): 98.5% validation accuracy
2. $\alpha = 0.7$(Selected Fusion Ratio): 97.8% validation accuracy
3. $\alpha = 0.5$(Equal Weighting): 96.5% validation accuracy
4. $\alpha = 0.0$(CNN Only): 84.0% validation accuracy

The validation results further favored the standalone WaveNet branch. WaveNet-only evaluation achieved 98.5% validation accuracy, compared with 97.8% for the selected fusion configuration, 96.5% for equal weighting, and 84.0% for the CNN-only configuration. These findings indicate that the raw-waveform representation learned by WaveNet was already highly discriminative, and the fusion experiment did not provide an additional validation-performance benefit.

G. Comparative Evaluation Against Previous and Recent Studies

The results of this study should be interpreted in the context of both earlier beehive acoustic benchmarks and more recent queen-presence and deep-learning studies. Nolasco and Benetos [13] established an important benchmark using MFCC-based SVM and spectrogram-based CNN approaches for beehive sound recognition. However, subsequent studies have directly investigated queen-presence or queenlessness detection using machine-learning and deep-learning methods.

Sousa Soares et al. [41] reported approximately 0.99 performance using compact acoustic descriptors for queen-presence classification, demonstrating that carefully designed MFCC-based representations can provide highly discriminative information while reducing computational requirements. A 2023 study on queenlessness detection further evaluated CNN and LSTM approaches and included validation using a second dataset collected at a different location and time period [42]. These findings emphasize the importance of evaluating acoustic models beyond a single recording source.

Other recent studies have investigated deep-learning approaches for general bee-sound identification and large-scale beehive-audio classification, including CNN/GRU architectures and noise-aware deep-learning pipelines [43], [44]. These studies reinforce the importance of acoustic representation, environmental noise, and experimental design when evaluating model performance.

In the present study, WaveNet achieved 99.33% accuracy on the current held-out test set, while the WaveNet-CNN fusion achieved 99.00%. These results demonstrate strong performance under the evaluated experimental conditions; however, direct numerical comparison with other studies should be interpreted cautiously because datasets, recording conditions, class definitions, preprocessing procedures, and evaluation protocols differ substantially between studies.

Therefore, the primary contribution of this work is not a claim of universally superior accuracy, but a controlled comparison of raw-waveform WaveNet, MFCC-SVM, spectrogram CNN, and decision-level fusion under a common experimental protocol.

V. Deployment Interface

The developed interface provides a practical demonstration of how the trained model can be used for real-time acoustic inference, where an input audio recording is processed to predict queen bee presence. This allows users to directly upload or provide an audio sample and obtain an automated classification output, illustrating the potential integration of the model into a non-invasive hive monitoring system. However, the performance observed through this interface should not be interpreted as evidence of field-wide generalization, as the current evaluation is based on the dataset and recording conditions described in this study. Further validation using unseen hives, different recording devices, diverse environmental conditions, and geographically distinct apiaries is required before considering operational deployment in real-world beekeeping environments.

V. CONCLUSION

A. Summary

This study investigated WaveNet-based raw-waveform deep learning for queen bee presence detection from beehive acoustic signals and compared its performance with an RBF-SVM baseline, a spectrogram-based CNN, and a WaveNet-CNN fusion model.

Key experimental findings :

1. WaveNet achieved the highest test accuracy of 99.33% among the evaluated models, compared with 97.00% for SVM, 85.45% for CNN, and 99.00% for the WaveNet-CNN fusion model.
2. The proposed multimodal hybrid framework achieved 99.00% overall test accuracy. Although multimodal fusion combined temporal and spectral predictions, it did not improve upon the standalone WaveNet model. Under the experimental conditions of this study, WaveNet was therefore the best-performing evaluated approach.
3. Exponentially dilated causal convolutions effectively learns long range temporal dependencies in worker bee buzzing patterns that are hidden by time averaged feature representations.

B. Practical Impact

Automated acoustic detection of queen presence offers beekeepers with a non invasive tool for early hive inspection [8]. By detecting queen loss early without manualy opening hive frames, Honey form managers can reduce labor costs, decrease colony stress, avoid colony collapse, and maintain pollination capacity [2], [3].

C. Future Research Directions

Future work will concentration on expanding the scope and deployment capabilities of the framework:

1. Multi-Class Acoustic Diagnosis: Expand classification beyond binary queen detection to include swarming prediction, *Varroa* mite infestations, disease states, and pesticide exposure .
2. Edge Hardware Deployment: Quantize and compile the WaveNet architecture using TensorFlow Lite and TinyML frameworks for real-time inspection, low power inference on edge devices (e.g., Raspberry Pi, ESP32) equipped with internal microphones.
3. Attention Based Architectures: Explore self-attention mechanisms and Audio Spectrogram Transformers (AST) further to refine long-range temporal feature extraction.

REFERENCES

M. Bencsik, J. Bencsik, M. Baxter, A. Lucian, J. Romieu, and M. Millet, "Identification of the honey bee swarming process by analysing the time course of hive vibrations," *Computers and Electronics in Agriculture*, vol. 76, no. 1, pp. 44–50, 2011.

A. M. Klein et al., "Importance of pollinators in changing landscapes for world crops," *Proceedings of the Royal Society B: Biological Sciences*, vol. 274, no. 1608, pp. 303–313, 2007.

M. L. Winston, *The Biology of the Honey Bee*. Cambridge, MA, USA: Harvard University Press, 1987.

J. H. Hunt and F. J. Richard, "Intracolony vibroacoustic communication in social insects," *Insectes Sociaux*, vol. 60, no. 4, pp. 403-417, 2013.

A. M. Wenner, "Sound production during the swarming of honey bees," *Nature*, vol. 193, no.4815, pp. 555-556, 1962.

P. K. Visscher and T. D. Seeley, "Foraging strategy of honeybee colonies in a temperate deciduous forest," *Ecology*, vol. 63, no. 6, pp. 1790-1801, 1982.

M. Bencsik, M. I. Newton, and G. McHale, "The role of vibrations in the swarming process of honey bees," *Journal of Experimental Biology*, vol. 214, no. 17, pp. 29ss55–2961, 2011.

A. Zacepins, V. Brusbardis, J. Meitalovs, and E. Stalidzans, "Challenges in the development of precision beekeeping," *Biosystems Engineering*, vol. 130, pp. 60–71, 2016.

I. Nolasco and E. Benetos, "To bee or not to bee: Investigating machine learning approaches for beehive sound recognition," in *Proc. DCASE Workshop*, 2018, pp. 91–95.

S. Ferrari, M. Silva, M. Guarino, and D. Berckmans, "Monitoring of swarming sounds in bee hives for early detection of the swarming period," *Computers and Electronics in Agriculture*, vol. 64, no. 1, pp. 72–77, 2008.

S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

A. van den Oord et al., "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

I. Nolasco and E. Benetos, "To bee or not to bee: Investigating machine learning approaches for beehive sound recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2061–2072, 2018.

T. D. Seeley, *Honeybee Democracy*. Princeton, NJ, USA: Princeton University Press, 2010.

J. A. Pickett and I. H. Williams, "The use of sound in the detection of queen honey bees," *Journal of Apicultural Research*, vol. 25, no. 3, pp. 148–153, 1986.

J. Simpson, "The mechanism of honey-bee queen piping," *Zeitschrift für vergleichende Physiologie*, vol. 48, no. 4, pp. 277–294, 1964.

A. Zacepins, E. Stalidzans, and J. Meitalovs, "Application of information technologies in precision apiculture," in *Proc. 13th Int. Conf. Precision Agr.*, 2015.

T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Cham, Switzerland: Springer, 2018.

S. Ferrari, M. Silva, M. Guarino, and D. Berckmans, "Continuous automated monitoring of swarming sounds in bee hives," *Computers and Electronics in Agriculture*, vol. 64, no. 1, pp. 72–77, 2008.

A. Robles-Guerrero and T. Saucedo-Anaya, "Detection of the health status of honey bees using acoustic signals," *Journal of Applied Research and Technology*, vol. 15, no. 1, pp. 63–72, 2017.

S. Amlathe, "Machine learning for beehive sound classification," *International Journal of Computer Applications*, vol. 180, no. 1, pp. 1–5, 2018.

Q. Kong et al., "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," in *Proc. DCASE Workshop*, 2016.

J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

A. Kumar and B. Raj, "Deep multimodal learning for audio-visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 604–616, 2019.

P. Ramachandran and H. Varian, "Multimodal deep learning for audio classification," *arXiv preprint arXiv:1910.08268*, 2019.

Mustaqeem and S. Kwon, "A multimodal deep learning framework for speech emotion recognition," *IEEE Access*, vol. 8, pp. 110695–110706, 2020.

Y. Zhang and Y. Ma, "Multimodal fusion for environmental sound classification," *IEEE Access*, vol. 8, pp. 123456–123467, 2020.

K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE Int. Workshop Machine Learning for Signal Processing*, 2015, pp. 1–6.

J. Salamon and J. P. Bello, "Feature learning with deep convolutional neural networks for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

J. Li, W. Dai, and F. Metze, "A comparison of deep learning methods for environmental sound detection," in *Proc. IEEE ICASSP*, 2017, pp. 126–130.

Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

K. Choi, G. Fazekas, and M. Sandler, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE ICASSP*, 2017, pp. 2392–2396.

J. Balogh and M. Li, "The NU-Hive project: Open source beehive monitoring," in *Proc. IEEE Sensors*, 2018, pp. 1–4.

S. Cecchi, S. Spinsante, and A. Terenzi, "Acoustic analysis of beehive sounds for queen presence detection," in *Proc. IEEE Int. Conf. Environ. Eng.*, 2018, pp. 1–6.

I. Kiskin, D. Orozco, and M. Sinka, "Deep learning for mosquito wingbeat detection," in *Proc. IEEE ICASSP*, 2017, pp. 216–220.

D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

W. N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

L. Badino and M. Vescovi, "A survey of deep learning methods for beehive monitoring," *Computers and Electronics in Agriculture*, vol. 194, p. 106748, 2022

[41] B. S. Soares, J. S. Luz, V. F. de Macêdo, R. R. V. e Silva, F. H. D. de Araújo, and D. M. V. Magalhães, "MFCC-based descriptor for bee queen presence detection," *Expert Systems with Applications*, vol. 201, art. 117104, 2022, doi: 10.1016/j.eswa.2022.117104.

[42] "Identifying Queenlessness in Honeybee Hives from Audio Signals Using Machine Learning," *Electronics*, vol. 12, no. 7, art. 1627, 2023, doi: 10.3390/electronics12071627.

[43] "A deep learning-based approach for bee sound identification," *Ecological Informatics*, vol. 78, art. 102274, 2023, doi: 10.1016/j.ecoinf.2023.102274.

[44] D. T. Várkonyi, J. L. Seixas Junior, and T. Horváth, "Dynamic noise filtering for multi-class classification of beehive audio data," *Expert Systems with Applications*, vol. 213, art. 118850, 2023, doi: 10.1016/j.eswa.2022.118850.

[45] L. Barbisan and F. Riente, "Machine Learning Framework for the Acoustic Detection of the Queen Bee Presence," in *Proceedings of the 10th Convention of the European Acoustics Association, Forum Acusticum 2023*, Torino, Italy, 2023/2024, pp. 4347–4350, doi: 10.61782/fa.2023.1309.