IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Leveraging Machine Learning Approaches to Decode Hive Sounds for Stress Prediction

SABA MUSTAFA[1], MAHSA MOHAGHEGH[2], IMAN ARDEKANI [3], and ABDOLHOSSEIN SARRAFZADEH[4]

[1]School of Engineering, Computing and Mathematical Sciences, Auckland University of Technology (AUT), Auckland 1010, New Zealand (saba.mustafa@autuni.ac.nz)
[2]School of Engineering, Computing and Mathematical Sciences, Auckland University of Technology (AUT), Auckland 1010, New Zealand (mahsa.mohaghegh@aut.ac.nz)
[3]The University of Notre Dame Australia (iman.ardekani@nd.edu.au)
[4]Center of Excellence in Cybersecurity Research, Education and Outreach (CREO) at NC A&T State University, Greensboro, NC, USA (hasarrafzadeh@ncat.edu)

Corresponding author: Mahsa Mohaghegh (e-mail: mahsa.mohaghegh@aut.ac.nz).

ABSTRACT Beekeeping plays a vital role in preserving ecosystems through pollination and increasing biodiversity. Effective monitoring of honeybee health and hive conditions is essential to balance bee populations and their environment. This study addresses the challenges of data scarcity and generalization in beehive health monitoring by introducing a semi-supervised learning model that employs a Transformer-based encoder-classifier for acoustic analysis of hive sounds. This research demonstrates the application of a Transformer-based architecture specifically tailored for bee bioacoustics and stress detection, integrating advanced feature extraction and fine-tuning techniques for this application. The main objective is to identify stress-related indicators from audio data collected via smart beehives. The proposed method utilizes a dataset of 5,336 labelled audio clips from diverse sources, including the NU-hive project and YouTube audio, to aid the learning process and enhance the classification accuracy for both labeled and unlabeled data. The audio features used in the analysis include Mel-frequency cepstral coefficients (MFCCs) and their delta and delta-delta variants, root mean square (RMS) energy, spectral centroid, and dominant frequency from Short-Time Fourier Transform (STFT). The Transformer-based encoder-classifier is implemented to classify bee behaviour within the hive as Normal, NoQueen, or Swarm, and to distinguish stressed from not stressed states. Evaluations indicate that the semi-supervised Transformer encoder-classifier achieves 99% accuracy on labeled data, with precision and recall values of 0.99 or higher for the Normal and NoQueen classes, and 0.96 for the Swarm class. Cluster validation produced a silhouette score of 0.47 and a Davies-Bouldin index of 0.57, indicating moderate cluster separability and compactness. The model was able to pseudo-label 94.7% of unlabeled data, validated against the nearest labelled neighbours. These results show the effectiveness of AI-driven beehive monitoring in supporting sustainable beekeeping practices and ecosystem conservation efforts.

INDEX TERMS Acoustic analysis, Beehive health monitoring, Honeybee colony stress detection, Honeybee health, Machine learning, Precision beekeeping, Semi-supervised learning, Smart beehives, Sustainable beekeeping, Transformer-Encoder architecture.

I. INTRODUCTION

HONEYBEES are important to pollination and reproduction in the ecosystem and, therefore, are indispensable to the environment [1]. However, the decline in bee populations has been observed in recent years, caused by habitat loss, pesticide exposure, climate change, and diseases, revealing the fragile nature and importance of this species [2]. The role of beekeepers is very significant in mitigating disease impacts and supporting the nutrition of bees, but they face challenges due to the remote locations of bee farms and difficulty in regular beehive monitoring [3]. To mitigate these issues, researchers and practitioners have been exploring non-invasive or minimally invasive techniques to become aware of the internal conditions of colonies without disturbing the hives [4], [5]. Among various indicators, the sounds produced by beehives have proven to be a key factor in evaluating colony health and behavior. Studies have shown that honeybees produce specific sounds when they experience stress from factors such as pest infections, airborne toxicants, swarming events, and missing queens [6]. Table 1 shows the list of stressors

and research using advanced AI techniques to investigate their impact on bee health and colony populations.

This research aims to develop an advanced method for predicting stress levels in bee colonies using acoustic features and state-of-the-art machine learning techniques. The implication of this study lies in its potential to change how bee colony health is monitored. The study aims to detect early signs of stress indicators within the hive that may lead to colony collapse. This non-invasive technique could provide timely information to beekeepers to take preventive actions to mitigate the impact of CCD and other threats to bee populations. By integrating a Transformer-based encoder architecture within a semi-supervised learning framework, this work addresses key challenges in data scarcity and generalization that have limited previous approaches. In doing so, it advances the field of precision beekeeping by providing a robust, scalable solution for real-time colony health assessment.

One of the most emerging methods of assessing hive health is performing audio analysis using machine learning (ML) techniques to make predictions based on collected audio data. Our literature review of recent publications in databases such as IEEE Xplore, SpringerLink, ScienceDirect, and PubMed indicates increasing research interest in bee health monitoring using acoustic and sensor-based techniques. In [7], [8], researchers have shown that their classification techniques achieved an accuracy close to 99% in detecting queen presence by extracting frequency-dependent coefficients from raw audio [33]. Most features employed in these analyses include Mel-frequency cepstral coefficients, Mel spectrograms, and short-time Fourier transforms. It has been observed that bees produce sounds with various frequencies depending on the colony's condition and activities. Normal buzzing typically ranges from 190-250 Hz, while stressed conditions such as queenlessness, swarming, or virgin queen emergence produce distinct frequency patterns. For instance, during swarming, bees exhibit frequencies between 100-250 Hz, while queenless colonies show patterns distinct from the 400 Hz pattern of healthy colonies [24].

Further advancements in artificial intelligence, particularly convolutional neural networks (CNN) and recurrent neural networks (RNN), have shown promising results in audio recognition tasks related to beehive monitoring. These approaches have demonstrated performance comparable to or surpassing traditional machine learning models in classifying beehive audio [32], [35].

A. RESEARCH OBJECTIVES

This study aims to enhance the emerging field of audio-based beehive health monitoring by investigating advanced machine-learning techniques for detecting stress in honeybee colonies. The research addresses two key questions:

- Can the integration of advanced audio feature extraction with Transformer-based models improve the clustering performance and classification reliability for stress state detection in bee colonies?
- Can a Transformer-based encoder accurately cluster and classify stress conditions in bee colonies using latent feature representations, and how does clustering validate its performance?

By exploring these questions, we aim to design more efficient and accurate techniques for detecting stressed bee behavior using acoustic analysis. This paper is structured as follows. Section 1 explains the importance of distinguishing stressed and non-stressed bees. Section 2 presents related work in bee acoustics and advanced ML-based bee sound detection. Section 3 discusses the proposed methodology, focusing on feature extraction and the implementation of a Transformer-enhanced encoder model for latent representation learning and classification. Section 4 presents clustering and classification results. Finally, Section 5 concludes the study.

II. RELATED WORK

Recent years have seen significant advancements in beehive monitoring and bee health assessment through the application of acoustic analysis and machine learning techniques. Researchers have explored various approaches across different domains to improve the overall understanding of bee behaviour, colony health, and environmental interactions. Figure 2 illustrates the distribution of research approaches in bee monitoring studies, distinguishing between pure and overlapping methodological contributions across 54 reviewed papers.This visualization highlights the prevalence of hybrid methodologies in the field and clarifies the distinct and combined contributions of each research approach.

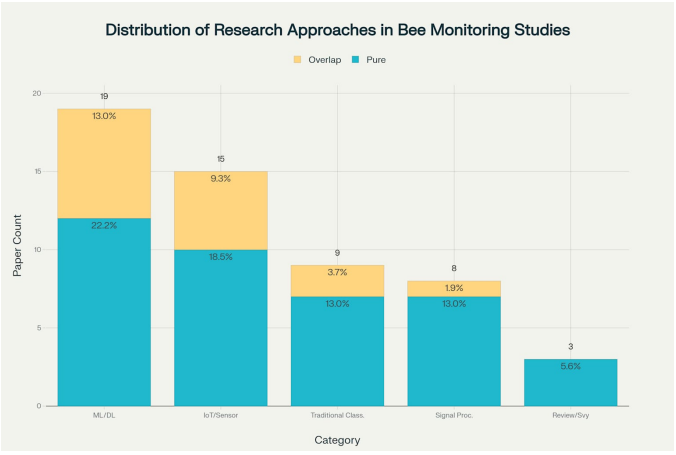


FIGURE 1. Research category distribution showing pure and overlapping methodological approaches across 54 bee monitoring studies

A. ACOUSTIC ANALYSIS AND MACHINE LEARNING TECHNIQUES

A prominent approach in bee acoustic monitoring involves the use of machine learning and deep learning techniques to analyze sound data from beehives. Several studies have employed Mel Frequency Cepstral Coefficients (MFCCs) as a primary feature extraction method for audio analysis [2],

TABLE 1. Types of Stressors and Studies using Acoustic Analysis

Stressor/Condition	Description	Relevant Studies
Queen Bee Presence/Absence	Detecting the presence or absence of a queen bee using acoustic signals	Early signal processing for queenless state identification [7], [8]. LSTM networks for queen-lessness detection [9]. Sound signal analysis of queenless colonies [2]. TinyML monitoring for queen detection [10]. Acoustic monitoring for queen presence [1].
Swarming Behavior	Predicting or detecting swarming events through acoustic analysis	Swarm activity acoustic classification [11]. Vibrational spectra for swarm prediction [12]. Biosensor signals for swarm prediction [13]. Audio features for swarming prediction [14]. Vibration monitoring for swarming [1].
General Colony Health	Assessing overall colony health and stress levels	Sound emission analysis [3]. Multi-sensor data mining [15]. Sensor-based health monitoring [16]. AI-based health evaluation [17]. Internal monitoring systems [18]. Multisensory health monitoring [19].
Toxicity Assessment	Detecting exposure to pesticides through acoustic analysis	Environmental pollution monitoring [20]. Air pollutant detection [21].
Insect Infestation	Identifying ant infestations using sound signals	Invasive insect recognition [22]. Varroa mite detection [23].
Winter Survivability	Predicting colony survival through winter using acoustic data	Multi-modal prediction of winter survival [24]. Winter monitoring system [25].
Drone Flight Detection	Detecting drone flights using audio signals	Audio-based detection [26], [27]. Selective feature analysis [28].
Bee Sound Identification	Classifying and identifying various bee sounds	Deep learning for sound identification [5]. VGGish embedding for sound classification [29]. Machine learning models [30]. CNN processing [31]. Spectrogram-based analysis [32]. Multi-class classification [33]. Found detector system [34].

[3], [5], [35]. MFCCs have proven effective in capturing the spectral characteristics of bee sounds, enabling researchers to distinguish between different colony states and behaviours. In [5], Truong introduced a novel deep learning approach called Mel-CNN-GRU, which combines Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs) for bee sound identification. This method demonstrated superior performance in distinguishing bee buzzing sounds from noise and cricket sounds, achieving a 1% improvement in accuracy over previous models. Similarly, in [35], Ruvinga employed Long Short-Term Memory (LSTM) networks and CNNs for identifying queenlessness in beehives, achieving test accuracy of 90.8% and 98.61%, respectively. The application of deep learning techniques has not been limited to sound classification. In [32], Borgianni explored the use of DenseNet121, ResNet50, and InceptionV3 models for analyzing spectrograms of beehive sounds. The study also highlighted the potential of federated learning approaches in beehive monitoring, which could address privacy concerns and enable collaborative model training across multiple apiaries. In [23], Voudiotis developed a deep-learning beehive monitoring system for early detection of the Varroa mite, a significant threat to bee colonies. This research demonstrates the potential of acoustic analysis combined with deep learning for pest detection and overall beehive health monitoring.

B. NOISE FILTERING AND FEATURE EXTRACTION

A significant challenge in the acoustic monitoring of beehives is dealing with background noise and extracting relevant features from audio data. In [33], Várkonyi addressed this issue by developing a dynamic noise filtering method for spectrograms, which outperformed existing baselines in the multi-class classification of beehive audio data. Their work emphasized the importance of effective noise filtering techniques in improving the performance of machine learning models for bee sound analysis. In addition to MFCCs, researchers have explored other feature extraction methods such as the Hilbert-Huang Transform (HHT) and Continuous Wavelet Transform (CWT) [3], [29]. These techniques offer alternative approaches to capturing time-frequency information from bee sounds and may provide complementary insights when combined with traditional MFCC-based methods. In [34], Kiromitis developed a bee sound detector that is easy to install, low-power, and low-cost. This system focuses on efficient noise filtering and feature extraction to monitor beehive conditions accurately.

C. MULTI-MODAL APPROACHES AND ENVIRONMENTAL MONITORING

Recent research has greatly emphasized the benefits of associating several sensory modalities for an extensive beehive monitoring system. In [24], [36], [37], the authors demonstrated the feasibility of combining audio data and humidity and temperature measurements to predict the winter survivability of honeybee hives. Their multimodal approach received an AUC-ROC score of 0.730, reinforcing the importance of using many sources of information to evaluate the health of beehives. The integration of environmental sensors with acoustic monitoring has also been explored by several researchers [10], [37]–[39]. These studies emphasize the importance of considering external factors such as temperature, humidity, and weather conditions when interpreting bee sounds and behaviour. The BeeLive platform developed by [38] exemplifies this approach, combining various sensor data with acoustic monitoring to provide comprehensive insights into hive health and performance. In [15], Braga proposed a method for mining combined data from in-hive sensors, weather information, and apiary inspections to forecast honeybee colony health status. This multi-modal approach demonstrates the potential for more accurate predictions by integrating various data sources. In [40], Hong presented a long-term and extensive monitoring system for bee colonies based on the Internet of Things (IoT). Their work showcases the potential of IoT technologies in beehive monitoring, allowing for continuous data collection and analysis. In [41], Bellino presented an integrated multi-sensor system for remote bee health monitoring. This research further emphasizes the importance of combining multiple sensor types for a more comprehensive understanding of hive conditions. In [42], Aydin and Aydin designed and implemented a smart beehive monitoring system using microservices in the context of IoT and open data. Their work demonstrates the potential of modern software architecture in creating scalable and flexible beehive monitoring solutions.

D. ECOACOUSTIC CODES AND COMMUNICATION

In addition to monitoring colony health, researchers have investigated the possibility of Ecoacoustic codes within beehives. In [6], Farina (2023) explored how honeybee buzzing patterns may integrate with external environmental cues, suggesting that complex acoustic patterns could govern nuanced forms of bee communication. This study applied algorithms such as the acoustic complexity index (ACI) and hierarchical K-means clustering to isolate emergent patterns, termed potential Ecoacoustic codes (PECs).

E. EMERGING TRENDS AND INNOVATIVE APPROACHES

Recent research has explored novel approaches to beehive monitoring. In [31], Sakova investigated beehive acoustic monitoring and processing using convolutional neural networks and machine learning. In [43], Kontogiannis developed a beehive smart detector device for critical conditions using edge device computations and deep learning inferences. In [26], [27], [44], Libal and Biernacki conducted several studies on audio-based bee classification, including MFCC selection by LASSO, MFCC-based sound classification, and a non-intrusive system for honeybee recognition based on audio signals and maximum likelihood classification by autoencoder. In [17], Liang proposed developing an AI-based integrated system for bee health evaluation, while in [45], Abdollahi introduced the UrBAN dataset for urban beehive acoustics and phenotyping. Many studies highlight challenges with beehive monitoring datasets, particularly the scarcity of open-access data, which limits research progress. Table 2 summarizes key datasets and studies that have utilized them for various beehive monitoring applications. Existing public datasets, such as Buzz, Nu-hive, and OSBH, are often small and focus narrowly on bee and queen detection. The MSPB dataset offers a multi-sensor approach with phenotypic measurements but lacks raw audio signals, providing only pre-processed data. Similarly, the UrBAN dataset addresses some limitations with over 2000 hours of raw audio collected over two years but still faces challenges in generalizing across diverse hives. A key issue is that models trained on individual datasets often fail to generalize to unseen hives due to variability in hive conditions. Efforts like the *BeeTogether* dataset aim to merge multiple datasets to increase diversity and standardize data formats for better model extrapolation. These challenges underscore the need for larger, more diverse datasets and standardized frameworks to improve the robustness of beehive monitoring systems.

III. MATERIALS AND METHODS

This section outlines the materials and methods for predicting bee behaviour using audio data collected from diverse beehives. The workflow is divided into two main phases: audio pre-processing and audio classification. In the pre-processing phase, relevant features are extracted from the audio signals. The classification phase employs a semi-supervised learning approach using a Transformer encoder-based deep learning model. This model is specifically designed to process sequences of extracted features, leveraging self-attention mechanisms to capture temporal patterns in bee acoustics and improve classification accuracy, even with limited labelled data.

Figure 2 shows the workflow that breaks down the research process into various steps. In the initial stage, feature extraction techniques are applied to labelled audio data to extract meaningful features. These features are then fed into a Transformer encoder-based model, which is trained on the labelled data. The model subsequently utilizes its knowledge to label the unlabeled data, classifying Bee Audio as either stressed or not stressed based on what it learned from the training data. Our thorough research process ensures the validity and reliability of our findings.

A. DATASET USED

This study utilizes two datasets: (1) labelled data from the NU-Hive Project (augmented with OSBP and YouTube audio

TABLE 2. Overview of Acoustic Analysis Techniques and Datasets in Beehive Monitoring Research

Technology/Method Used	Sub-Category	Dataset Used	References
Signal Processing			
Spectrograms	Sound Analysis	NU-Hive dataset (2 hives, 5 days)	[3], [32], [46]
MFCC Analysis	Sound Analysis	MSPB dataset (53 hives, 365 days)	[26], [27], [29], [47]
Wavelet Analysis	Sound Analysis	OSBH dataset (6 hives, 2 days)	[4], [48]
HHT Analysis	Sound Analysis	BUZZ dataset (6 hives, 109 days)	[12], [21]
Machine Learning			
SVM	Audio/Classification	NU-Hive + OSBH combined	[33], [49]
Random Forests	Audio/Classification	BeeTogether dataset	[5], [35], [48]
k-NN	Audio/Classification	TBON processed dataset	[34], [50]
Deep Learning			
CNN	Audio Processing	SBCM dataset (4 hives)	[31], [51]
LSTM	Audio Processing	UrBAN dataset (10 hives, 2000+ hours)	[9], [24], [45]
VGGish	Audio Processing	MSPB dataset	[29]

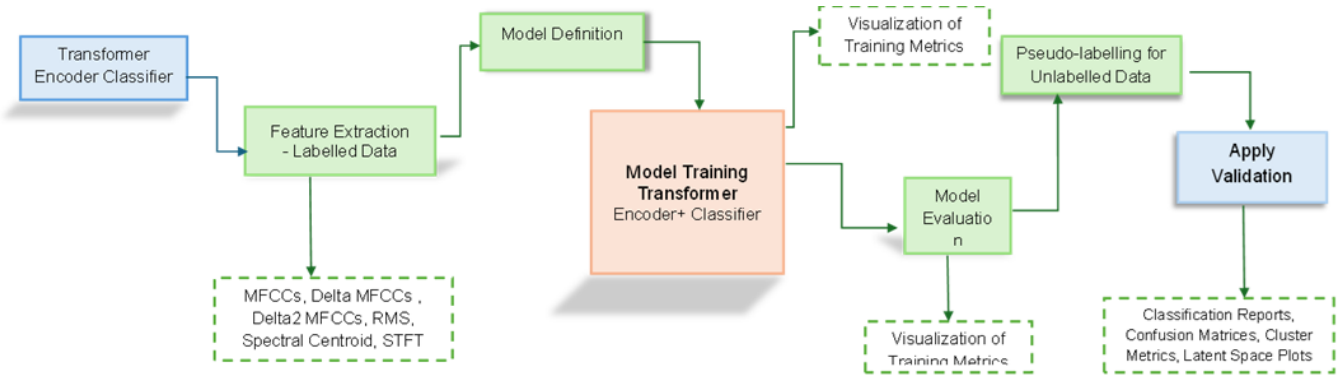


FIGURE 2. Illustration of the workflow of the proposed Transformer Encoder Classifier framework for bee behaviour prediction.

clips) and (2) unlabelled data from the BeeAudio dataset. Regardless of their original length or source, all audio recordings were re-segmented into 5-second clips to increase the number of training samples and capture fine-grained bee activity, since shorter homogeneous segments have been shown to improve deep learning performance on bee sounds [52], [53].

a: Labelled Dataset

The first dataset comprises labelled data from the *NU-Hive Project*, which provides annotated audio recordings of bee sounds collected from different beehives [4], [54]. This dataset includes recordings from two hives in two different states: queenless and normal/active. The dataset contains 576 audio files, equally divided between these two states, representing two hives over two days each. Originally 10 minutes long, these recordings were segmented into 1-minute clips to enhance the feature extraction process [46], [54]. To improve the *NU-Hive* dataset, especially concerning swarming behaviour, short audio clips ranging from 5 to 15 seconds were obtained from multiple YouTube videos [55], [56]. Furthermore, a 10-minute swarm video from the *Open Source Beehives Project (OSBP)* Dataset was incorporated into this labelled dataset [57].

b: Unlabelled Dataset

The second dataset used in this study is unlabelled data from the *BeeAudio* Dataset [58]. This dataset represents the largest open-access, sound-focused honeybee dataset with multimodal data. It was created to support computational approaches addressing bee population decline, such as developing machine learning algorithms for remote and instant detection of hive health status via sound data. The *BeeAudio* Dataset was collected using a custom IoT device combining an ESP32 Wi-Fi module, an INMP441 microphone module, and a BME280 temperature/humidity sensor. It contains 7,100 samples of 60-second audio clips from European Honeybee hives in California, making it particularly suitable for identifying bee behaviour and timely detection of stressed bees.

Due to computational resource constraints, this study utilized a subset of 558 BeeAudio clips. These 60-second samples were further divided into non-overlapping 5-second segments for feature extraction and model training. Integrating these diverse datasets establishes a strong basis for the study, enabling in-depth analysis of bee behaviour across various situations and supporting the development of sophisticated beehive health monitoring systems [46]. While the datasets include recordings from multiple hive types and geographic locations, explicit evaluation of the model on entirely new

environments or under different noise conditions was not conducted in this work. Future research will focus on domain adaptation and cross-location validation to further enhance the robustness and applicability of the model in varied real-world settings.

c: Ethical and Licensing Considerations

All YouTube audio clips used in this study were sourced from publicly available videos in accordance with YouTube's Terms of Service, used solely for non-commercial academic research, and attributed where possible. No copyrighted material was redistributed.

B. DATA PRE-PROCESSING AND FEATURE EXTRACTION

All data processing was implemented in Python (v3.11.11) on Google Colab, leveraging PyTorch's `DataLoader` (with a custom `collate_fn`) for efficient batching, padding, and shuffling of variable-length audio sequences. [59].

1) Data Segmentation

Since the machine learning models are intended to perform multi-class classification, using short 5-second audio clips is adequate to attain an accuracy of over 98% [50]. The audio files in the datasets were segmented to meet the short-length audio clip requirement. Long audio recordings are divided into fixed-length chunks for analysis using audio segmentation to facilitate a machine-learning approach. Using the Python library Librosa, the audio files are loaded from the input directory, and each audio file's duration and sampling rate are determined. Subsequently, the audio signal is divided into non-overlapping segments with a predetermined duration of 5 seconds. The audio segmentation is achieved by calculating the number of samples per segment and slicing the signal accordingly. For consistency, only segments of the desired length are maintained. Audio segmentation and feature extraction were performed in-memory without writing intermediate audio files, enabling efficient processing and training.

2) Feature Extraction

One of the most critical aspects of this research is sound analysis, specifically audio processing. Since machine learning models cannot directly comprehend unprocessed audio input, significant features must be extracted. Sound data must be preprocessed before implementing Machine Learning or deep learning approaches. Due to the multi-dimensional structure of audio data, which is characterized by various frequencies that change over time, preprocessing approaches must be used for efficient feature extraction. Feature extraction helps to observe different signal characteristics. Typically, audio signals undergo pre-processing to provide spectral characteristics, which can graphically depict changes in energy across time and frequency. Consequently, these extracted audio features are directly incorporated into a Transformer encoder-based model, enabling it to learn complex temporal and spectral patterns in bee sounds and accurately classify colony states and stress indicators. The effectiveness of multimodal and multi-feature approaches has been well established in speech emotion recognition, where combining diverse audio and contextual features leads to more robust classification systems [60], [61].This motivates the integration of multiple acoustic features in our bee sound analysis pipeline. In the context of bee acoustics, several feature extraction methods have been identified and utilized in recent studies:

- **Mel-frequency cepstral coefficients (MFCCs)**: MFCCs have been widely applied in bee sound analysis. In [62], Nolasco demonstrated their effectiveness in identifying beehive states.
- **Delta and Delta-Delta MFCCs**: First- and second-order derivatives of MFCCs, capturing both velocity and acceleration of cepstral changes, as used in this study. In [4], Terenzi employed delta MFCCs in their comparative study of feature extraction methods for honeybee activity classification.
- **Root Mean Square (RMS)**: RMS provides information about the overall energy of the audio signal. [15], [63] Braga utilized RMS in their method for forecasting honeybee colony health status.
- **Spectral Centroid**: This feature represents the "centre of mass" of the spectrum. In [24], Zhu incorporated spectral centroid in their multi-modal approach to predicting honeybee hive winter survivability.
- **Dominant Frequency**: In [12], Ramsey used dominant frequency analysis in predicting swarming in honeybee colonies.

The selection of RMS energy, spectral centroid, and delta MFCCs was guided by their demonstrated effectiveness in previous bee acoustic studies and their ability to capture key aspects of bee sound relevant to stress detection. While features such as harmonic-to-noise ratio (HNR) and zero-crossing rate (ZCR) have been used in general bioacoustic research, they were not included in this study to prioritize features with the strongest empirical support for bee colony state classification. A formal feature importance analysis (e.g., SHAP values or permutation importance) was not conducted in this work but is identified as a valuable direction for future research to further optimize and interpret the feature set.

The researchers proposed that honeybee sounds can provide valuable insights regarding overall colony health and behaviour [2], [3], [30], [63]. The selected features—MFCCs, delta/delta-delta MFCCs, spectral centroid, dominant frequency, and RMS energy—are chosen for their documented relevance to bee stress indicators. For instance, dominant frequency and spectral centroid directly capture the frequency shifts characteristic of queen piping (330–430 Hz) and swarming (100–500 Hz) events, while MFCCs encode the spectral and temporal patterns associated with both normal and stressed states. RMS energy provides a measure of signal intensity, which can increase during agitation or defensive behaviors. Although this study does not include a direct feature-to-event correlation analysis, the literature

and the frequency ranges summarized in Table 3 support the biological relevance of these features for detecting stress-related acoustic events in bee colonies. Future work will focus on more granular mapping between extracted features and specific stress events, such as piping or wingbeat anomalies.

The Table 3 demonstrates that bee acoustic signals traverse a broad frequency band depending on the colony's condition and behaviour. Normal activities typically fall within the 100-1000 Hz range, while stress or threat responses can extend to much higher frequencies. Notably, acoustic patterns are greatly influenced by the queen's presence or absence, especially in the lower frequency ranges. However, as shown in Table 3, different states have overlapping frequencies, making it difficult to categorize bee behaviour using only frequency bands. It can be challenging to determine the precise state of the hive because a frequency of 400 Hz, for instance, could indicate either a healthy colony or queen piping. Analyzing audio characteristics other than frequency, such as amplitude, modulation, and temporal patterns, is essential to overcoming this constraint.

This study implemented advanced feature extraction methods such as MFCCs, delta-delta MFCCs, RMS energy, spectral centroid, and dominant frequency to capture the diverse acoustic characteristics of bee sounds. By integrating these features into a Transformer encoder-based model, the approach enables highly precise multi-class classification of bee behaviours and colony states within the hive. This multi-faceted methodology not only improves the model's ability to distinguish between overlapping frequency patterns, but also provides a more comprehensive assessment of colony stress levels and overall health, as reflected in the achieved 99% classification accuracy

3) Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is a foundational tool for analyzing the time-varying frequency content of bee sounds in acoustic monitoring. By segmenting the audio signal into overlapping frames and applying the Fourier Transform to each frame, STFT enables the capture of dynamic changes in bee acoustic patterns, such as those associated with queen piping, swarming, or stress events [3], [14], [32], [66]. The STFT provides constant absolute bandwidth analysis for identifying harmonic components and offers consistent resolution in two-dimensional representation, regardless of the actual frequency. The mathematical expression for the STFT is given by:

$$S(f,\tau) = \sum_{t=0}^{N-1} x(t)\omega(t-\tau)e^{-j2\pi ft} \tag{1}$$

where $x(t)$ is the analyzed signal, and $\omega(t)$ is the window function centered at time τ. In this study, the STFT is used as an intermediate step for extracting spectral features, such as MFCCs, their dynamic derivatives, and dominant frequency, rather than as a direct input to the deep learning model. This approach enables the model to capture both the static and dynamic spectral properties of bee sounds.

The STFT is implemented using the Python library Librosa, which efficiently computes the frequency content of each frame over time. These spectral representations are further processed for feature extraction, providing essential input for downstream machine learning and deep learning classification tasks. By leveraging STFT-based features, the model achieves robust performance in distinguishing between different bee behaviours and colony states, as reflected in the high classification accuracy reported in this study.

4) Mel Frequency Cepstral Coefficients (MFCC)

Mel-frequency cepstral coefficients, or MFCCs, are a valuable technique for assessing complicated bee sounds and behaviours in the context of bee acoustic monitoring. MFCCs transform audio signals into perceptually meaningful features by mapping the spectrum onto the Mel scale and then applying cepstral analysis, which effectively captures the spectral envelope and timbral characteristics of bee acoustics [26], [27]. This process emulates aspects of human auditory perception and has been widely adopted for tasks such as bee and queen presence detection, swarming identification, and hive health assessment [2], [5], [6], [8] In this study, MFCCs, along with their first-order (delta) and second-order (delta-delta) derivatives, are extracted for each frame of the audio signal to capture both static and dynamic spectral properties. This approach aligns with recent research in the field of bee acoustics and precision beekeeping. For instance, [4] Terenzi conducted a comprehensive comparison of feature extraction methods for sound-based classification of honey bee activity, highlighting the effectiveness of MFCCs. The computation of MFCCs is typically performed using specialized libraries, such as `librosa.feature.mfcc`, with 40 coefficients (`n_mfcc=40`) to capture the spectral characteristics of the audio signal based on human auditory perception. The process involves several steps:

- The audio signal is transformed into a spectrogram using the Short-Time Fourier Transform (STFT).
- The spectrogram is mapped to the Mel scale, emphasizing lower frequencies in alignment with the human auditory system.
- The logarithm of the Mel spectrogram is decorrelated using the Discrete Cosine Transform (DCT).
- Delta and delta-delta coefficients are calculated to capture temporal changes and acceleration in the spectral features [7], [10].

In our pipeline, each 5-s segment is framed with `n_fft=2048` and `hop_length=512` , this process yields a temporal sequence of 40 MFCCs, 40 delta MFCCs, and 40 delta-delta MFCCs per frame, resulting in a rich representation of both the spectral content and its variation over time. These features are not averaged but retained as sequences, enabling the deep learning model to leverage both the static and dynamic aspects of bee acoustics for robust classification.

TABLE 3. Key Acoustic Signals Associated with Different Colony Behaviours and Stress Levels, Along with Their Corresponding Frequency Ranges

Condition/Behavior	Frequency Range	Stress Indicator	Notes	Reference
Normal Worker Buzz	255 ± 35 Hz	Calm	Basic hive sound during normal activity	[2]
Normal Colony Activity	100-1000 Hz	Calm	General range with respective harmonics	[3]
Healthy Colony	300, 410, 500 Hz	Calm	Characteristic pattern of queenright colony	[30]
Queen Piping	330-430 Hz	Calm	Fundamental frequency with little modulation	[64]
Swarming Signals	100-200 Hz to 500+ Hz	High stress	Worker piping prior to swarming with high variability	[65]
Tooting (emerged virgin queens)	200-550 Hz	Neutral	Typically around 400 Hz	-
Hissing (Defense)	300-3600 Hz	Increased stress	Broad band noise during distress or threats	[65]
Hornet Attack Response	5000 Hz with harmonics up to 15-16 kHz	Stress	Guard bees produce distinct hissing sounds	[3]
Pre-swarming Activity	110 Hz increasing to 300 Hz	Increased stress	Progressive increase in amplitude and frequency	[64]
Pipping (potential swarm indicator)	340-450 Hz	Increased stress	Challenge signal from queen	-
Queenless State	<1080 Hz	Stressed	Large fluctuations in lower frequencies	[2]

This approach yields a frame-wise feature sequence that the deep learning model uses for accurate multi-class classification of bee behaviours and colony states.

5) Other Feature Extraction Techniques

This research extracted a comprehensive set of spectral and energy-based features from audio signals to analyze their acoustic properties deeply. The Root Mean Square (RMS) Energy was computed using the `librosa.feature.rms` with a frame length of 2048 and hop length of 512, RMS measures the instantaneous energy or loudness of the signal, providing insight into the intensity of bee activity. In addition to energy, the Spectral Centroid was extracted to capture the brightness of the sound, representing the "centre of mass" of the spectrum. The centroid was computed for each signal frame, weighted by the magnitude of the spectrum using the `librosa.feature.spectral_centroid` function. The default configurations were used with a Fourier Transform window size (`n_fft` of 2048) and a hop length of 512 samples to ensure high spectral resolution. This feature is commonly used for timbre analysis and characterizing the spectral balance of an audio signal. Finally, the Dominant Frequency was identified to determine the most prominent tonal component of the signal. The `librosa.fft_frequencies` function was used to compute the frequency values corresponding to the Short-Time Fourier Transform (STFT) bins, with the default FFT size set to 2048. The energy across time frames was summed for each frequency bin using `np.sum(stft, axis=1)`, and the bin with the highest energy was selected using `np.argmax`. The frequency matched to this bin was specified as the dominant frequency. This attribute detects the signal's most significant pitch or tonal component. These properties create a strong signal representation, including its tonal qualities, spectral balance, and intensity, for further research analysis.

6) Feature Vector Representation

Finally, a comprehensive feature vector was constructed for each frame by concatenating diverse acoustic features to provide a robust, time-resolved representation of every audio segment. This feature vector sequence integrates 40 Mel-frequency cepstral coefficients (MFCCs), 40 first-order derivatives (delta MFCCs), and 40 second-order derivatives (delta-delta MFCCs), capturing both the static and dynamic spectral characteristics of the signal. These are augmented with Root Mean Square (RMS) energy to represent signal intensity, Spectral Centroid to reflect spectral balance or brightness, and Dominant Frequency to identify the most prominent tonal component. Mathematically, at each time frame t, a 123-dimensional feature vector $\mathbf{f}_t$ is constructed as:

$$\begin{aligned}\mathbf{f}_t = \big[&\mathrm{MFCC}_1(t), \ldots, \mathrm{MFCC}_{40}(t),\\ &\Delta\mathrm{MFCC}_1(t), \ldots, \Delta\mathrm{MFCC}_{40}(t),\\ &\Delta^2\mathrm{MFCC}_1(t), \ldots, \Delta^2\mathrm{MFCC}_{40}(t),\\ &\mathrm{RMS}(t), \mathrm{SC}(t), \mathrm{DF}(t)\big]\end{aligned} \tag{2}$$

Where:

- $\mathrm{MFCC}_i(t)$ is the i-th Mel-Frequency Cepstral Coefficient at time frame t
- $\Delta\mathrm{MFCC}_i(t)$ is the first-order temporal derivative (delta)
- $\Delta^2\mathrm{MFCC}_i(t)$ is the second-order temporal derivative (delta-delta)
- $\mathrm{RMS}(t)$ is the Root Mean Square energy
- $\mathrm{SC}(t)$ is the Spectral Centroid
- $\mathrm{DF}(t)$ is the Dominant Frequency

The complete feature matrix $\mathbf{F} \in \mathbb{R}^{T\times 123}$ for an audio clip with T time frames is:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1^\top \\ \mathbf{f}_2^\top \\ \vdots \\ \mathbf{f}_T^\top \end{bmatrix} \in \mathbb{R}^{T \times 123} \tag{3}$$

The resulting sequence of 123-dimensional vectors for each audio segment forms the input to the deep learning model. This approach ensures that the extracted features collectively capture the critical spectral, temporal, and tonal aspects of bee acoustics, making them suitable for detailed analysis and robust classification in this study.

C. DEEP LEARNING ALGORITHMS USED

The research employed a distinct methodology: a semi-supervised learning approach utilizing a Transformer encoder-based model. A small set of labelled audio data from diverse sources, including the NU-Hive project, was used. The model was trained in a semi-supervised manner, where labelled data directly guided classification, while the Transformer model generated pseudo-labels for the unlabelled data, enabling the model to iteratively improve its performance on both sets. This integrated approach allowed effective learning from limited labelled examples, resulting in high multi-class classification accuracy for bee behaviour and colony state prediction.

1) Dataloader

This research implemented data loading using PyTorch's data handling utilities to prepare and load the dataset for model training efficiently.First, we converted each audio segment's frame-wise feature sequence into a PyTorch tensor with floating-point values, and did the same for the labels, using integer types. We then bundled these into a single dataset structure and used PyTorch's DataLoader to handle batching, shuffling, and, with a custom function, padding variable-length sequences. This setup made it straightforward and efficient to feed the data into our Transformer-based model for training and analysis. [59]. This DataLoader implementation offers several benefits for model training: it enables automated batch size management, efficient shuffling to reduce bias, and correct handling of variable-length acoustic sequences through padding and masking. This benefit shows promising results in recent studies using large-scale datasets like MSPB [36] and UrBAN [45], where efficient data handling is crucial for model performance. This approach is particularly valuable for processing bee acoustic data, as it enables efficient handling of the high-dimensional MFCC features extracted from continuous hive monitoring. The DataLoader's batching capabilities and shuffling method help avoid biases in the model learning process by randomly presenting various bee behaviours and conditions. This methodology has been used in recent studies of precision beekeeping [29], [30] that emphasize the need for data handling, regardless of its size and complexity, for real-time colony monitoring and accurate behaviour classification. The approach used in this research focuses on the optimal utilization of computational resources while maintaining the temporal relationships crucial for analyzing bee acoustic patterns, supporting both research objectives and practical applications in apiary management.

2) Model Creation & Training: Transformer Encoder-Based Model

This study introduces a Transformer-based architecture designed for robust feature extraction and sequence modelling in bee acoustics analysis. The model leverages deep neural encoders and Transformer modules to process high-dimensional, frame-wise acoustic features and capture complex temporal dependencies relevant to bee behaviour classification. Transformers, introduced by [67], have revolutionized sequence modelling tasks across various domains. Originally designed for natural language processing, Transformers have shown remarkable adaptability to other fields, including audio processing. Their self-attention mechanism allows for capturing long-range dependencies in sequential data, making them particularly suitable for analyzing the temporal patterns in bee sounds [68].

The Transformer-based architectur illustrated in Figure 3 integrates an encoder with a Transformer module, leveraging the strengths of both dimensionality reduction and sequence modeling for bee acoustic analysis. The workflow begins with input audio features represented as a 2D tensor $X \in \mathbb{R}^{B \times D}$, where B denotes the batch size and $D = 128$ represents the input dimensionality. The encoder processes this input to reduce its dimensionality and extract a compact 64-dimensional latent representation. This is achieved through two fully connected (dense) layers with configurations of 128 and 64 neurons, respectively. Mathematically, the encoder transformation can be expressed as:

$$h_1 = \text{ReLU}(W_1 x + b_1), \quad h_2 = \text{ReLU}(W_2 h_1 + b_2) \tag{4}$$

Where $x \in \mathbb{R}^{128}$ is the input feature vector, $W_1 \in \mathbb{R}^{128 \times 128}$ and $W_2 \in \mathbb{R}^{64 \times 128}$ are the weight matrices of the first and second fully connected layers, respectively, $b_1 \in \mathbb{R}^{128}$ and $b_2 \in \mathbb{R}^{64}$ are the corresponding bias vectors. The function $\text{ReLU}(\cdot)$ denotes the Rectified Linear Unit activation function, defined as:

$$\text{ReLU}(z) = \max(0, z) \tag{5}$$

which introduces non-linearity and enhances the model's ability to learn complex patterns. The resulting 64-dimensional latent representation $h_2 \in \mathbb{R}^{64}$ is then passed to the Transformer module, which operates on a 3D tensor with dimensions (B, S, D), where S is the sequence length and $D = 64$ is the input dimension. Positional encoding is added to retain temporal information. The Transformer module is designed to capture global contextual information in sequential or time-series data. It consists of a 64-dimensional latent space, four attention heads, and two Transformer layers, balancing model complexity and computational efficiency. The

self-attention mechanism within the Transformer computes attention scores as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Where Q, K, and V are the query, key, and value matrices, respectively, derived from the input h_2 through learned linear transformations, and d_k is the dimensionality of the keys (typically $d_k = 64$ in this case). The softmax function ensures that the attention weights sum to one.

The outputs of the multi-head attention layers are concatenated and passed through a feed-forward network with layer normalization and residual connections, as per the standard Transformer architecture. This process is repeated across two Transformer layers to refine the latent representation. For classification, the output representations are averaged across the time dimension, and a final fully connected layer maps this aggregated representation to the desired number of output classes, producing predictions suitable for bee behavior classification. This Transformer-based architecture integrates dimensionality reduction, self-attention, and feature aggregation for classification to efficiently process sequential audio data. The combination of dense layers, self-attention, and ReLU activation ensures an optimal balance between computational efficiency and representational power, making it well-suited for complex audio signal analysis.

3) Model Training

The model training process involves optimizing the performance of the Transformer-based classifier using labelled audio features. This process ensures that the model learns to classify bee states based on extracted features while retaining its ability to generalize to unseen data. The training process uses the Cross-Entropy Loss function to minimize the classification error between the predicted and true labels. This ensures accurate classification of audio signals into their respective classes (e.g., "Normal," "NoQueen," "Swarm"). The training begins with a forward pass, where input features are passed through the encoder to produce a compressed latent shape representation $(\text{batch size}, \text{sequence length},, 64)$. Positional encoding is added, and the sequence is processed by the Transformer encoder, which captures complex dependencies among features. The output is then averaged across the time dimension, and a final fully connected layer predicts class probabilities. The model's performance is optimized using the Cross-Entropy Loss function, defined as:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{7}$$

where C is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i. This loss is minimized during training to improve classification accuracy. In the backward pass, gradients of the loss concerning model parameters are computed using backpropagation. Following the rule, these gradients are then used to update the model parameters via the Adam optimizer with a learning rate of 0.001:

$$\phi \leftarrow \phi - \eta \nabla_\phi L \tag{8}$$

where ϕ represents the model parameters, η is the learning rate, and L is the loss. This iterative process continues until the model converges, ensuring effective learning and generalization.

IV. RESULTS

In this section, the study will explain the empirical analysis of the research. All experimentation was performed using Google COLAB. Below are the details related to the evaluation of the Semi-Supervised Learning approach using Transformer-based model to predict bees' behaviour.

A. MODEL EVALUATION & VISUALIZATION

The evaluation and visualization process for the Transformer-based classifier includes assessing its classification performance, visualizing latent feature space, and analyzing clustering metrics for both labelled and unlabelled data as presented in Table 4. In this study, experiments were performed

TABLE 4. Summary of pseudo-labeling and clustering-based validation methods commonly used in machine learning and semi-supervised learning.

Method	Purpose	Reliability
Latent Space Visualization	Check cluster formation for pseudo-labels	High (Visual Inspection)
Nearest Labeled Neighbor	Compare pseudo-labels to closest labeled data	High (Quantitative Check)
Cluster Validation Metrics	Measure cluster quality	Medium (Depends on Data)
Confidence Thresholding	Identify low-confidence pseudo-labels	High (Practical Check)
Feature Analysis by Label	Analyze feature separation across pseudo-labels	Medium (Feature-Specific)

using features extracted from honey bee audio data, including MFCCs, delta-MFCCs, delta2-MFCCs, spectral centroid, RMS, STFT, and dominant frequency, all computed using the `librosa` Python package [69]. The PyTorch library was used to build and train the custom Transformer-based architecture. For optimal model performance, key hyperparameters such as learning rate, number of Transformer layers, and batch size were systematically tuned. Grid Search, implemented via the scikit-learn library in Python [38], exhaustively evaluated all combinations of these hyperparameters. Table 5 summarizes the specific hyperparameters used in this study, the ranges explored, and the optimal values selected for the final model. For model evaluation, a train/test split was used for supervised classification. Performance was assessed using classification reports (precision, recall, F1-score, accuracy), confusion matrices, and visualizations of training/validation loss and accuracy. Latent feature space was visualized using t-SNE and PCA, and clustering quality

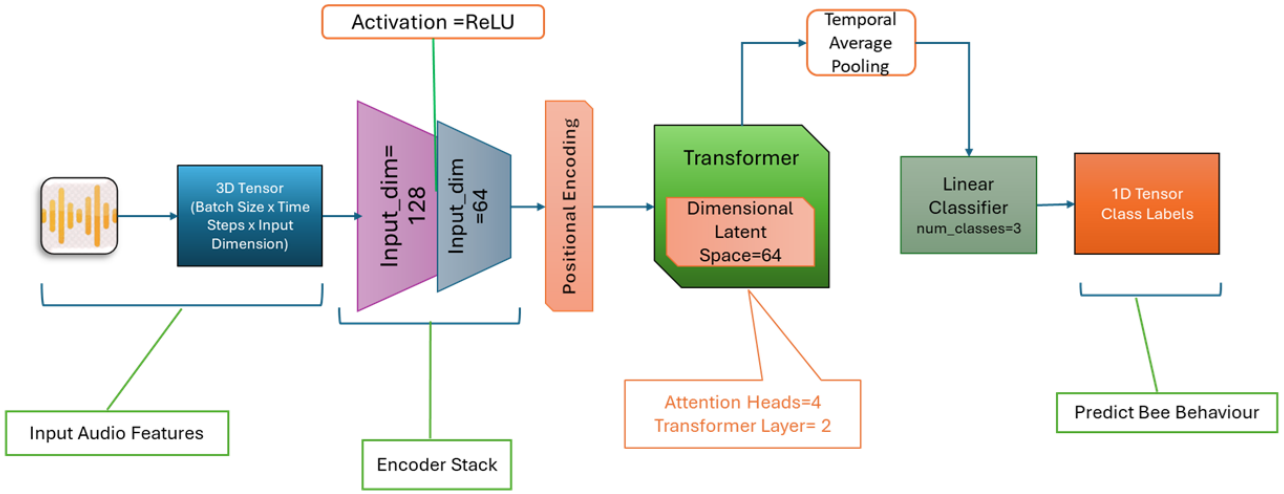


FIGURE 3. **Architecture of the proposed Transformer-based bee behavior classification model.The input audio features are passed through a linear encoder stack, followed by positional encoding and a Transformer encoder with two layers and four attention heads. Temporal average pooling is applied across time steps, and a linear classifier predicts the final bee behavior class labels.**

TABLE 5. **Hyperparameter Tuning Summary**

Hyperparameter	Value Range	Optimal Value
Learning Rate	[0.0001, 0.001, 0.01, 0.1]	0.001
Transformer Layers	[1, 2, 3]	2
Batch Size	[8, 16, 32]	8

was evaluated using Silhouette Score and Davies-Bouldin Index. Pseudo-labels for unlabelled data were validated using nearest neighbor comparisons and clustering metrics, as summarized in Table 4.

1) Evaluation of Model with Labelled Data

The performance of the Transformer-based model is evaluated on the test dataset using multiple metrics. A confusion matrix compares true and predicted labels, offering insights into the model's class-wise performance. Each element in the matrix represents the counts of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). This matrix is visualized using a heatmap for easy identification of misclassification patterns.

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{9}$$

The confusion matrix in Figure 4 represents the final evaluation of the model's performance on the labeled test set, showing how well the model predicts the true classes for three categories: Normal, NoQueen, and Swarm. The diagonal values represent correctly classified samples, while off-diagonal values indicate misclassifications. The model achieved an overall accuracy of 99%, correctly classifying 1060 out of 1068 samples. Class-wise performance shows that the model performs exceptionally well for the Normal and NoQueen classes, with precision and recall values above 0.99. The Swarm class, while maintaining a perfect precision of 1.00, exhibits a slightly lower recall of 0.93, suggesting that a small proportion of "Swarm" samples were misclassified as

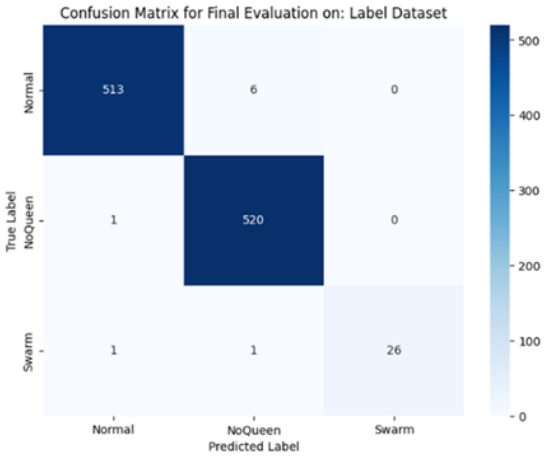


FIGURE 4. **Confusion Matrix for Label Data Classification**

other classes. These misclassifications, reflected in the off-diagonal elements, account for a few samples being predicted as "Normal" or "NoQueen."

Overall, the metrics derived from the confusion matrix, including a weighted F1-score of 0.99 and a macro F1-score of 0.98, highlight the model's reliability and strong generalization capabilities across all three classes. In addition, a classification report provides detailed metrics such as precision, recall, and F1-score for each class. Precision measures the proportion of true positive predictions among all predicted positives, while recall evaluates the proportion of true positives out of all actual positives. The F1-score combines these two metrics to offer a balanced assessment. These metrics are visualized as bar plots, offering an intuitive understanding of the model's performance. Finally, predictions for each test sample are saved to a CSV file, including the file name, true label, predicted label, and stress condition (stressed/not stressed) for post-hoc analyses.

The classification report demonstrates the efficacy of the Transformer-based model in identifying stress conditions in bee sounds across three classes: Normal, NoQueen, and Swarm. The model achieves an overall accuracy of 99% with a weighted F1-score of 0.99, showcasing its robustness. Class-wise analysis highlights near-perfect performance for the Normal and NoQueen classes, with precision and recall values above 0.99. For the Swarm class, although the precision remains 1.00, the recall is 0.93, indicating the model occasionally misclassifies true "Swarm" samples as other classes. The macro average F1-score of 0.98 confirms balanced performance across all classes, while the weighted average ensures that class imbalance does not impact the overall evaluation. These metrics affirm the model's reliability for stress detection in bee colonies, particularly for critical tasks like distinguishing between NoQueen and Swarm behaviors.

Figure 5 illustrates the training and validation accuracy of the Transformer-based model over 30 epochs. The training accuracy (blue line) steadily increases, reaching near-perfect accuracy (1.0) after approximately 20 epochs. Similarly, the validation accuracy (green line) follows a similar trend, plateauing at a high value (0.98–1.0) with minimal fluctuations. This alignment between training and validation accuracy demonstrates the model's ability to generalize well to unseen data, indicating minimal overfitting. The minor fluctuations in validation accuracy observed in the middle epochs are attributed to the inherent variability in the validation dataset. Overall, the results confirm the effectiveness of the model in learning meaningful patterns from the data and applying them to unseen samples. Figure 6 shows the training and validation loss trends over 30 epochs. The training loss (blue line) decreases sharply during the initial epochs, indicating rapid learning. The validation loss (orange line) closely tracks the training loss, demonstrating strong generalization to unseen data. Both losses plateau at very low values (0.05–0.1), indicating minimal overfitting and robust learning. Figure 7 illustrates the 2D latent space representation of labeled data samples generated by the encoder of the Transformer-based model. Each point represents a data sample, color-coded according to its true cluster label: Normal (purple), NoQueen (teal), and Swarm (yellow). The three distinct clusters demonstrate the model's ability to learn separable and compact representations for each class. The compactness of the clusters reflects the model's capability to encode intra-class similarities, while the well-separated clusters indicate effective learning of discriminative features. This visualization in confirms that the model can effectively distinguish between the three classes, supporting robust classification and generalization to unseen data.

In addition to the three-class classification (Normal, NoQueen, Swarm), we evaluated the model's performance by collapsing these categories into a binary scheme: "Not Stressed" (Normal) and "Stressed" (NoQueen, Swarm). Figure 8, the bar chart below, illustrates the distribution of pre-

TABLE 6. Classification Report for Label Data

Class	Precision	Recall	F1-Score	Support
Normal	1.00	0.99	0.99	519
NoQueen	0.99	1.00	0.99	521
Swarm	1.00	0.93	0.96	28
Overall Accuracy	-	-	0.99	1068
Macro Avg	0.99	0.97	0.98	1068
Weighted Avg	0.99	0.99	0.99	1068

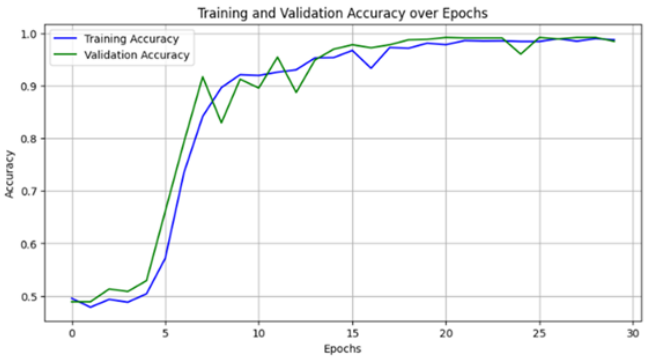


FIGURE 5. Training and Validation Accuracy for Label Data

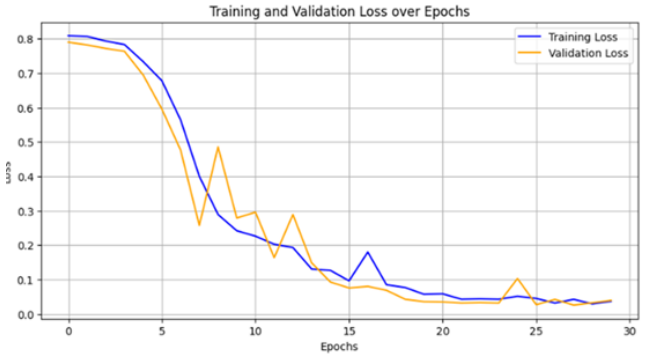


FIGURE 6. Training and Validation Loss for Label Data

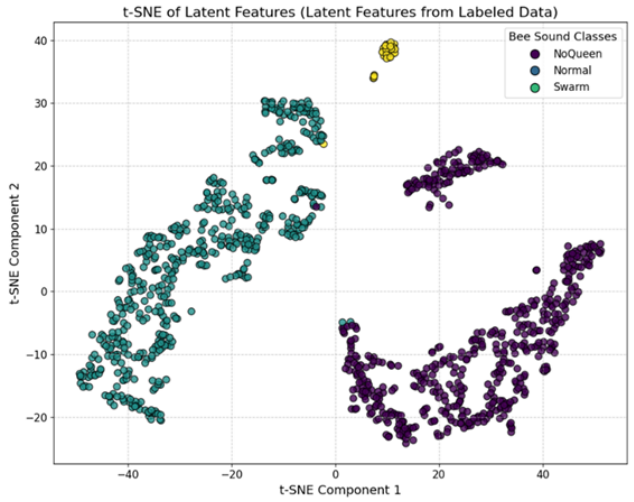
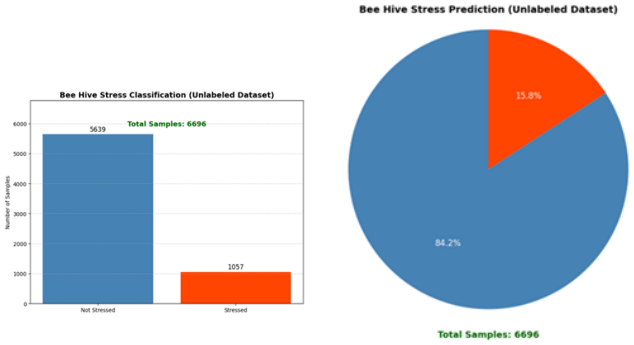


FIGURE 7. **Training and Validation Loss for Label Data**



(a) Bar chart showing absolute counts.

(b) Pie chart showing percentage distribution.

FIGURE 8. **Binary classification results on the pseudo-labeled unlabeled dataset. Subfigure (a) shows the number of samples predicted as "Stressed" and "Not Stressed," while subfigure (b) visualizes their percentage breakdown.**

dictions, with 5,639 samples classified as "Not Stressed" and 1,057 as "Stressed," out of a total of 6,696 samples. The accompanying pie chart shows that 84.2% of samples were predicted as "Not Stressed" and 15.8% as "Stressed." This binary analysis demonstrates that the model is capable of robustly identifying stress conditions relevant for practical hive management, providing beekeepers with a clear and actionable summary of hive health. Presenting results in this format aligns the model's output with real-world decision-making needs and supports timely intervention for colony well-being.

Overall, the results and all visual representations show that the Transformer-based encoder successfully achieves high performance on the labeled data, with excellent accuracy, precision, recall, and F1-score across all classes.

2) Evaluation of Model with Unlabeled Data

The evaluation of pseudo-labeling performance was conducted on previously unlabeled bee audio samples using the trained Transformer-based model. For each unlabeled sample, a pseudo-label was assigned based on the model's prediction. To validate the quality of these pseudo-labels, we compared them to the nearest labeled samples in the latent space using cosine similarity, following established semi-supervised learning practice. Figure 9 shows the confusion

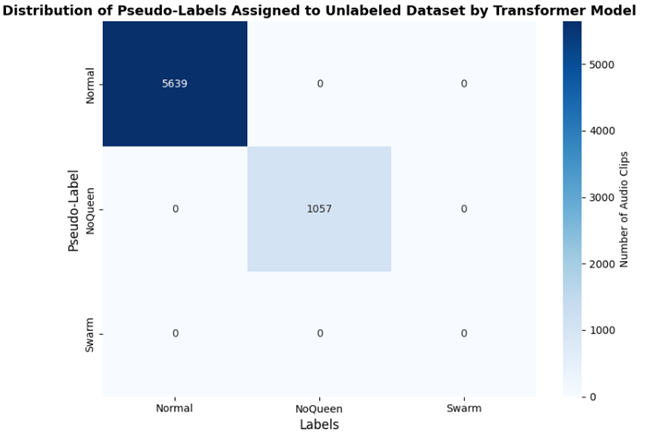


FIGURE 9. **Confusion matrix showing the distribution of pseudo-labels (Normal, NoQueen, Swarm) assigned by the trained Transformer model to the unlabeled datase.Each entry indicates the number of audio clips classified into each category.**

matrix for the pseudo-labeled dataset. The matrix summarizes the number of audio clips in the unlabeled dataset that were assigned to each class (Normal, NoQueen, Swarm) by the trained Transformer model. Specifically, 5,639 samples were labeled as "Normal," 1,057 as "NoQueen," and no samples were assigned to the "Swarm" class. This figure provides an overview of the predicted distribution of hive states in the unlabeled data, as determined by the model's pseudo-labeling.

Figure 10 presents the t-SNE visualization of the latent features extracted from the unlabeled data. Each point represents a sample, colored by its assigned pseudo-label ("Normal" in yellow, "NoQueen" in purple). The visualization reveals two well-separated clusters corresponding to the two classes, confirming that the model learned discriminative latent representations even for data not seen during supervised training. The clear separation of clusters indicates confident pseudo-labeling and robust feature extraction. These results demonstrate the effectiveness of the model's latent space in capturing the underlying structure of the data and ensuring high-quality pseudo-labeling. Notably, no pseudo-labels were assigned to the "Swarm" class in the unlabeled dataset (see Figure 9 and Figure 10). This indicates that either the unlabeled data did not contain any samples representative of the "Swarm" class, or the model was not able to confidently identify this class among the unlabeled samples. As a result, both the confusion matrix and t-SNE plot display only "Normal" and "NoQueen"

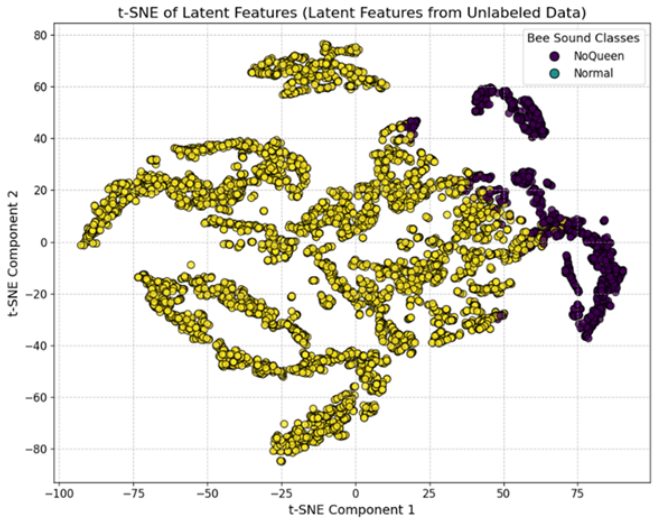


FIGURE 10. Latent Space Representation of Unlabeled Data

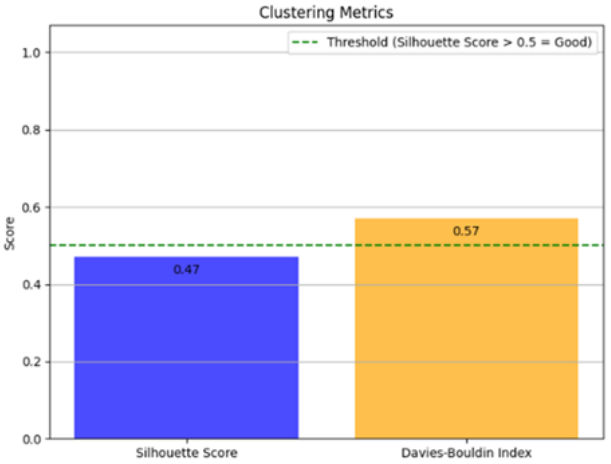


FIGURE 11. Clustering Metrics for Pseudo-Labeling Evaluation

categories. The visualizations confirm the effectiveness of the pseudo-labeling process, with the vast majority of samples assigned to clusters that align with the expected data structure. The compactness and clear separation of the "Normal" and "NoQueen" clusters in the t-SNE plot reflect the model's ability to encode intra-class similarities and learn discriminative features.

Overall, these results highlight the reliability of the Transformer-based encoder in distinguishing between the present classes and demonstrate the potential of this semi-supervised approach for robust classification and generalization in real-world bee sound monitoring applications. Pseudo-labels for unlabeled bee audio were validated by comparing each pseudo-labeled sample to its nearest labeled neighbor in the model's latent space using cosine similarity. No human expert review was performed; incorporating expert validation is suggested for future work.

3) Evaluation of Pseudo-Labeling on Unlabeled Data

In this section, the results of three evaluation methods—Clustering Metrics Visualization, Latent Space Visualization, and Nearest Labeled Neighbor—are discussed to verify that the pseudo-labeling is optimal and the model is able to label the dataset with high accuracy. The clustering quality of the pseudo-labels assigned to the unlabeled data was first assessed using Silhouette Score and Davies-Bouldin Index. As shown in Figure 11, the pseudo-label clusters achieved a Silhouette Score of 0.47 and a Davies-Bouldin Index of 0.57. According to the threshold indicated in the plot, a Silhouette Score above 0.5 is considered good, and a lower Davies-Bouldin Index indicates more compact and well-separated clusters. While the Silhouette Score is just below the "good" threshold, these values still indicate moderate to strong cluster compactness and separation in the latent space learned by the model. This confusion matrix evaluates the agreement between the pseudo-labels assigned to unlabeled data and the true labels derived from the nearest labelled neighbours in the latent space. The rows represent the labels of the nearest neighbours (ground truth), while the columns represent the pseudo-labels assigned by the model. Figure 12 shows the

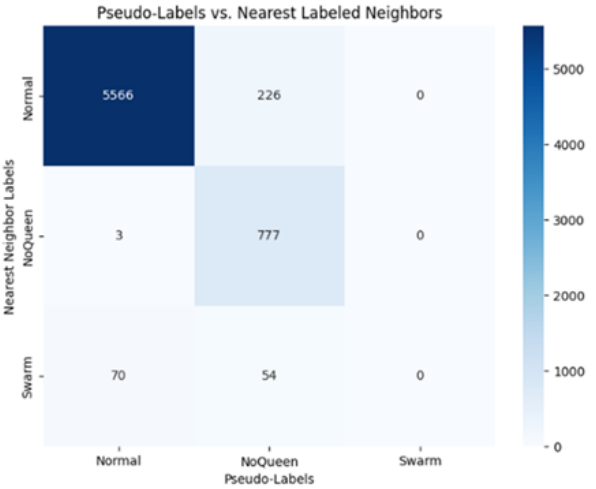


FIGURE 12. Confusion Matrix Comparing Pseudo-Labels to Nearest Neighbors

confusion matrix comparing the pseudo-labels assigned by the model to the labels of their nearest labeled neighbors in the latent space. The majority of samples are located on the diagonal, indicating strong agreement between the model's pseudo-labels and the structure learned from labeled data. Specifically, 5,566 samples were consistently labeled as "Normal" by both the model and their nearest neighbors, and 777 as "NoQueen." However, there is some overlap: 226 samples labeled as "Normal" by the model were closest to "NoQueen" in the latent space, and 70 "Swarm" samples were assigned the "Normal" pseudo-label. No samples were assigned the "Swarm" pseudo-label, reflecting either a lack of "Swarm" instances in the unlabeled data or model uncer-

tainty for this class. These findings validate the robustness of the Transformer-based encoder model in assigning pseudo-labels to the majority of samples. However, the results also highlight challenges in handling minority classes and overlapping feature distributions, suggesting avenues for further improvement in latent space representation and class-specific feature extraction.

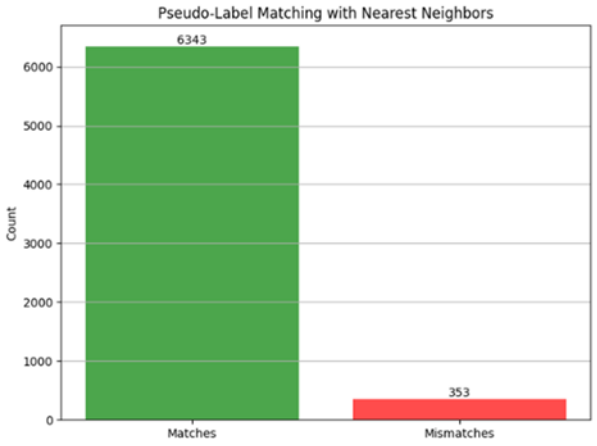


FIGURE 13. **Performance of Pseudo-Labeling vs. Nearest Neighbors**

Figure 13 visualizes the overall pseudo-label matching performance. Out of 6,696 unlabeled samples, 6,343 (94.7%) were matched with their nearest labeled neighbor, while 353 (5.3%) were mismatches. This high match rate demonstrates the model's strong generalization and its ability to transfer learned representations to previously unseen data. The small proportion of mismatches is largely attributable to overlapping feature distributions between the "Normal" and "NoQueen" classes, as indicated by the confusion matrix. These findings suggest that while the pseudo-labeling process is robust, additional refinement of the latent space representation or augmentation of training data could further improve accuracy.

Figure 14 presents the KMeans clustering of the latent representations (extracted from the model) projected into three principal components using PCA. The plot reveals three visually distinct clusters, with the majority of samples grouped into two large clusters corresponding to the "Normal" (yellow) and "NoQueen" (teal) classes, and a smaller, more diffuse cluster likely associated with "Swarm" or outlier samples. The clustering quality was quantified using silhouette and Davies-Bouldin scores: KMeans achieved a silhouette score of 0.38 and a Davies-Bouldin index of 0.91, indicating moderate cluster compactness and separation. These results are consistent with the confusion matrix and nearest neighbor analysis, confirming that the model's latent space effectively encodes class structure but also highlighting some overlap between classes.

To directly compare the clustering quality of pseudo-label assignments with classical clustering algorithms, Table 7

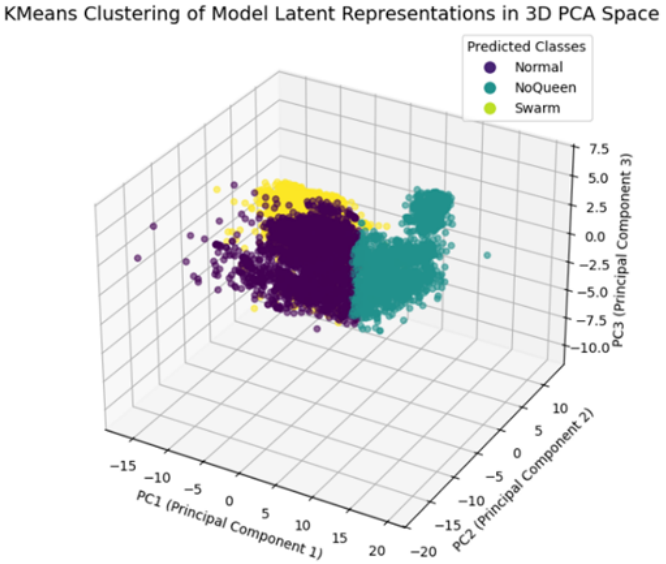


FIGURE 14. **KMeans clustering of latent representations extracted from the model, projected onto the first three principal components using PCA.**

summarizes the Silhouette Score and Davies-Bouldin Index for Pseudo-Labels, Agglomerative Clustering, and KMeans:

TABLE 7. **Clustering Quality Metrics for Different Methods**

Method	Silhouette Score	Davies-Bouldin Index
Pseudo-Labels	0.47	0.57
Agglomerative	0.387	0.866
KMeans	0.381	0.908

To assess the clustering quality of the latent space, we compared the Silhouette Score and Davies-Bouldin Index for clusters formed by pseudo-labels (assigned by the model) with those formed by standard unsupervised clustering algorithms (Agglomerative, KMeans). As shown in Table 7, the pseudo-label clusters achieved the highest Silhouette Score (0.47) and lowest Davies-Bouldin Index (0.57), indicating more compact and better-separated clusters than those produced by unsupervised methods. This demonstrates that the model's latent space is highly structured and clustering-friendly, supporting the effectiveness of the pseudo-labeling process. Overall, these results demonstrate that the Transformer-based model produces high-quality pseudo-labels for the majority of unlabeled data, with robust latent space organization and effective clustering. The absence of samples assigned the "Swarm" pseudo-label suggests either a scarcity of "Swarm" instances in the unlabeled set or a challenge for the model in confidently identifying this minority class. Future work could focus on further improving class separation and minority class detection through targeted data augmentation or advanced feature engineering. The comparative clustering analysis further confirms the superiority of the model's learned representations over classical clustering approaches, supporting its suitability for semi-supervised bee sound classification.

V. DISCUSSION

The focus of this study was to evaluate a Transformer-based encoder-classifier architecture, marking the first application of such a model specifically tailored for bee bioacoustics. This novel approach leverages the strengths of Transformers combined with an encoder for feature extraction and dimensionality reduction. Unlike previous studies that relied solely on supervised learning methods, this research adopts a semi-supervised learning framework, addressing critical challenges such as data scarcity and limited generalization. The methodology included both supervised classification on labeled data and pseudo-labeling of unlabeled data, followed by comprehensive validation using clustering and nearest-neighbor analyses. The training and validation results demonstrate the model's strong ability to generalize to labeled data. The Transformer-based model achieved an overall accuracy of 99% on the labeled test set, with precision and recall metrics of 0.99 or higher for the Normal and NoQueen classes, and 0.96 for the Swarm class (see Table 6). The confusion matrix revealed that most misclassifications occurred in the Swarm class, which had limited representation compared to the other classes. This is consistent with the observed overlap in latent space, where Swarm samples were more scattered and sometimes overlapped with Normal or NoQueen clusters. The pseudo-labeling process was evaluated using both clustering metrics and latent space visualizations. The Silhouette Score (0.47) and Davies-Bouldin Index (0.57) for the pseudo-label clusters indicate moderate cluster compactness and separation—outperforming traditional clustering algorithms such as Agglomerative (Silhouette: 0.39, DBI: 0.87) and KMeans (Silhouette: 0.38, DBI: 0.91) (see Table 7). This demonstrates that the model's learned latent space is more structured and discriminative than what is achieved by unsupervised clustering alone. Latent space visualizations further confirmed that the Normal and NoQueen classes formed well-separated clusters, while Swarm samples remained more diffuse and overlapped with other classes. Although the Silhouette Score for pseudo-label clusters is moderate (0.47), the high agreement rate (94.7%) between pseudo-labels and their nearest labeled neighbors demonstrates strong practical reliability for pseudo-label assignment. This indicates that, despite some cluster overlap, the majority of pseudo-labels remain accurate and trustworthy for downstream analysis and real-world hive monitoring. Alternative clustering methods, including Agglomerative and KMeans, were also evaluated and found to produce lower clustering quality than the Transformer-based approach, confirming the robustness of the model's latent space for pseudo-labeling. The moderate Davies-Bouldin Index (0.57) further supports the model's ability to distinguish stressed from non-stressed conditions with high reliability in practice. While some overlap between similar classes may occur, this clustering quality is sufficient for timely intervention and early warning in most scenarios, and ongoing data expansion can further improve reliability. The model consistently performed best on the well-represented Normal and NoQueen classes, both in labeled and pseudo-labeled data. The Swarm class, however, remained challenging due to its limited representation and overlapping acoustic characteristics. This suggests that increasing the diversity and quantity of Swarm samples, or enhancing feature extraction techniques, could further improve the model's performance for this class. The clustering metrics and visualizations were instrumental in understanding the model's limitations and strengths. While the latent space projections showed meaningful feature learning, the overlapping regions and scattered clusters for the Swarm class highlight opportunities for improvement. These could involve incorporating additional acoustic features, domain-specific transformations, or even hybrid approaches that combine audio and environmental data for stress detection. While the Transformer-based classifier demonstrated strong generalization and effective pseudo-labeling for the majority of data, challenges remain in separating overlapping classes and improving minority class (Swarm) detection. In our future work, we plan to address the challenge of minority class representation, particularly for the Swarm category, by implementing targeted strategies. We will explore data augmentation methods such as pitch shifting, time stretching, and noise injection to increase the diversity of Swarm samples. We also intend to augment the dataset for underrepresented classes, investigate advanced feature engineering, and incorporate additional domain knowledge—such as environmental or behavioral data—to further enhance model robustness and generalization across all bee colony states. Additionally, meta-heuristic or evolutionary optimization methods could be explored for automated feature selection, as demonstrated in recent speech emotion recognition studies [70]. Although the datasets used in this study include recordings from a variety of hive types and geographic locations (NU-Hive project, Open Source Beehives Project, BeeAudio from California), explicit evaluation on entirely new environments—such as different hive architectures, regions, or challenging noise conditions—was not conducted. The diversity of the current dataset likely contributes to the observed robustness and generalization on held-out data. However, the absence of formal domain adaptation or cross-location validation means the model's performance on truly unseen environments remains untested. Addressing this limitation is an important direction for future research, where domain adaptation techniques and targeted cross-location validation will be explored to ensure reliable performance in varied real-world settings.

VI. CONCLUSIONS

This study demonstrates the significant potential of advanced machine learning techniques in improving honeybee colony health monitoring. The research successfully addresses the challenges of data scarcity and generalization through a semi-supervised learning approach utilizing a Transformer-based encoder-classifier. Key achievements include:

- High classification accuracy of 99% in detecting bee stress states using a combination of MFCCs, delta and delta-delta MFCCs, RMS energy, spectral centroid, and dominant frequency features.

- Effective integration of advanced audio feature extraction with Transformer-based models, resulting in improved clustering quality (Silhouette Score: 0.47, Davies-Bouldin Index: 0.57) and reliable classification.
- Exceptional performance in classifying Normal and No-Queen classes, with precision and recall values of 0.99 or higher.
- Strong precision (1.00) and recall (0.93) for the Swarm class, despite challenges related to minority class representation and some overlap with other classes.

The study contributes significantly to audio-based beehive health monitoring, offering a reliable method for early stress detection in bee colonies. This research supports sustainable beekeeping practices and ecosystem conservation efforts by providing beekeepers with an efficient tool to monitor and maintain healthy colonies. Furthermore, this study aligns with several of New Zealand's Sustainable Development Goals, including protecting biodiversity (SDG 15), supporting food security (SDG 2), demonstrating technological innovation in agriculture (SDG 9), and contributing to climate resilience (SDG 13). By positioning New Zealand at the forefront of sustainable apiculture, this research reinforces the country's reputation for environmental stewardship and innovative agricultural practices [71].

ACKNOWLEDGMENT

The authors thank Auckland University of Technology (AUT) for their support and the generous Summer Scholarship, which was instrumental in the success of their research. They also acknowledge the use of Perplexity AI to enhance the language and clarity of this manuscript. All scientific ideas, designs, results, and conclusions are solely those of the authors.

REFERENCES

[1] C. Uthoff, M. N. Homsi, and M. V. Bergen, "Acoustic and vibration monitoring of honeybee colonies for beekeeping-relevant aspects of presence of queen bee and swarming," *Comput. Electron. Agric.*, vol. 205, p. 107589, Feb. 2023.

[2] D. K. et al., "Decoding the behavior of a queenless colony using sound signals," *Biology*, vol. 12, no. 11, p. 1392, Oct. 2023.

[3] A. Terenzi, S. Cecchi, and S. Spinsante, "On the importance of the sound emitted by honey bee hives," *Vet. Sci.*, vol. 7, no. 4, p. 168, Oct. 2020.

[4] A. Terenzi, N. Ortolani, I. Nolasco, E. Benetos, and S. Cecchi, "Comparison of feature extraction methods for sound-based classification of honey bee activity," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 112–122, 2022.

[5] T. H. Truong, H. D. Nguyen, T. Q. A. Mai, H. L. Nguyen, T. N. M. Dang, and T.-T.-H. Phan, "A deep learning-based approach for bee sound identification," *Ecol. Inform.*, vol. 78, p. 102274, Dec. 2023.

[6] A. Farina, "Discovering ecoacoustic codes in beehives: First evidence and perspectives," *Biosystems*, vol. 234, p. 105041, Dec. 2023.

[7] D. Howard, O. Duran, G. Hunter, and K. Stebel, "Signal processing the acoustics of honeybees (apis mellifera) to identify the queenless state in hives," in *SPRING CONFERENCE ACOUSTICS 2013*. Institute of Acoustics, Oct. 2023.

[8] G. Hunter, D. Howard, S. Gavreau, O. Duran, and R. Busquets, "Processing of multi-modal environmental signals recorded from a smart beehive," in *ACOUSTICS 2019*. Institute of Acoustics, May 2023.

[9] S. Ruvinga, G. J. A. Hunter, O. Duran, and J.-C. Nebel, "Use of lstm networks to identify 'queenlessness' in honeybee hives from audio signals," in *2021 17th International Conference on Intelligent Environments (IE)*. IEEE, Jun. 2021, pp. 1–4.

[10] A. D. Simone, L. Barbisan, G. Turvani, and F. Riente, "Advancing beekeeping: Iot and tinyml for queen bee monitoring using audio signals," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–9, 2024.

[11] A. Zgank, "Bee swarm activity acoustic classification for an iot-based farm service," *Sensors*, vol. 20, no. 1, p. 21, Dec. 2019.

[12] M.-T. R. et al., "The prediction of swarming in honeybee colonies using vibrational spectra," *Sci. Rep.*, vol. 10, no. 1, p. 9798, Jun. 2020.

[13] K. Iqbal, B. Alabdullah, N. A. Mudawi, A. Algarni, A. Jalal, and J. Park, "Empirical analysis of honeybees acoustics as biosensors signals for swarm prediction in beehives," *IEEE Access*, vol. 12, pp. 148 405–148 421, 2024.

[14] D. T. Várkonyi, D. T. Bányai, and A. R. Várkonyi-Kóczy, "Investigating traditional machine learning models and the utility of audio features for lightweight swarming prediction in beehives," *Acta Polytech. Hung.*, vol. 21, no. 10, pp. 283–299, 2024.

[15] A. R. Braga, D. G. Gomes, R. Rogers, E. E. Hassler, B. M. Freitas, and J. A. Cazier, "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies," *Comput. Electron. Agric.*, vol. 169, p. 105161, Feb. 2020.

[16] A. Zaman and A. Dorin, "A framework for better sensor-based beehive health monitoring," *Comput. Electron. Agric.*, vol. 210, p. 107906, Jul. 2023.

[17] A. Liang, "Developing an ai-based integrated system for bee health evaluation," *arXiv*, Jan. 18 2024.

[18] M. Micheli, G. Papa, I. Negri, M. Lancini, C. Nuzzi, and S. Pasinetti, "Sensorizing a beehive: A study on potential embedded solutions for internal contactless monitoring of bees activity," *Sensors*, vol. 24, no. 16, p. 5270, Aug. 2024.

[19] I. Rigakis, I. Potamitis, N.-A. Tatlas, G. Psirofonia, E. Tzagaraki, and E. Alissandrakis, "A low-cost, low-power, multisensory device and multivariable time series prediction for beehive health monitoring," *Sensors*, vol. 23, no. 3, p. 1407, Jan. 2023.

[20] N. P. et al., "Continuous monitoring of beehives' sound for environmental pollution control," *Ecol. Eng.*, vol. 90, pp. 326–330, May 2016.

[21] Y. Zhao, G. Deng, L. Zhang, N. Di, X. Jiang, and Z. Li, "Based investigate of beehive sound to detect air pollutants by machine learning," *Ecol. Inform.*, vol. 61, p. 101246, Mar. 2021.

[22] A. Nasir, M. O. Ullah, and M. H. Yousaf, "Ai in apiculture: A novel framework for recognition of invasive insects under unconstrained flying conditions for smart beehives," *Eng. Appl. Artif. Intell.*, vol. 119, p. 105784, Mar. 2023.

[23] G. Voudiotis, A. Moraiti, and S. Kontogiannis, "Deep learning beehive monitoring system for early detection of the varroa mite," *Signals*, vol. 3, no. 3, pp. 506–523, Jul. 2022.

[24] Y. Z. et al., "Early prediction of honeybee hive winter survivability using multi-modal sensor data," in *2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*. Pisa, Italy: IEEE, Nov. 2023, pp. 657–662.

[25] F. A. Arribas and M. R. Hortelano, "An internet of living things based device for a better understanding of the state of the honey bee population in the hive during the winter months," *Comput. Electron. Agric.*, vol. 212, p. 108026, Sep. 2023.

[26] U. Libal and P. Biernacki, "Mfcc selection by lasso for honey bee classification," *Appl. Sci.*, vol. 14, no. 2, p. 913, Jan. 2024.

[27] ——, "Mfcc-based sound classification of honey bees," *Int. J. Electron. Telecommun.*, pp. 849–849, Nov. 2024.

[28] F. Rustam, M. Z. Sharif, W. Aljedaani, E. Lee, and I. Ashraf, "Bee detection in bee hives using selective features from acoustic data," *Multimed. Tools Appl.*, vol. 83, no. 8, pp. 23 269–23 296, Aug. 2023.

[29] N. Di, M. Z. Sharif, Z. Hu, R. Xue, and B. Yu, "Applicability of vggish embedding in bee colony monitoring: Comparison with mfcc in colony sound classification," *PeerJ*, vol. 11, p. e14696, Jan. 2023.

[30] A. Robles-Guerrero, T. Saucedo-Anaya, C. A. Guerrero-Mendez, S. Gómez-Jiménez, and D. J. Navarro-Solís, "Comparative study of machine learning models for bee colony acoustic pattern classification on low computational resources," *Sensors*, vol. 23, no. 1, p. 460, Jan. 2023.

[31] M. Sakova, P. Jurik, P. Galajda, and M. Sokol, "Bee hive acoustic monitoring and processing using convolutional neural network and machine learning," in *2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, Apr. 2024, pp. 1–5.

[32] L. Borgianni, M. S. Ahmed, D. Adami, and S. Giordano, "Spectrogram based bee sound analysis with dnns: a step toward federated learning approach," in *2023 4th International Symposium on the Internet of Sounds*. IEEE, Oct. 2023, pp. 1–8.

[33] D. T. Várkonyi, J. L. Seixas, and T. Horváth, "Dynamic noise filtering for multi-class classification of beehive audio data," *Expert Systems with Applications*, vol. 213, p. 118850, Mar. 2023.

[34] D. I. Kiromitis, C. V. Bellos, K. A. Stefanou, G. S. Stergios, T. Katsantas, and S. Kontogiannis, "Bee sound detector: An easy-to-install, low-power, low-cost beehive conditions monitoring system," *Electronics*, vol. 11, no. 19, p. 3152, Sep. 2022.

[35] S. Ruvinga, G. Hunter, O. Duran, and J.-C. Nebel, "Identifying queenlessness in honeybee hives from audio signals using machine learning," *Electronics*, vol. 12, no. 7, p. 1627, Mar. 2023.

[36] Y. Z. et al., "Mspb: a longitudinal multi-sensor dataset with phenotypic trait measurements from honey bees," Nov. 2023.

[37] M. A. et al., "Recent developments on precision beekeeping: A systematic literature review," *Journal of Agricultural and Food Research*, vol. 14, p. 100726, Dec. 2023.

[38] A. S. Hamza, R. Tashakkori, B. Underwood, W. O'Brien, and C. Campell, "Beelive: The iot platform of bee-mon monitoring and alerting system for beehives," *Smart Agricultural Technology*, vol. 6, p. 100331, Dec. 2023.

[39] R. Tashakkori, A. S. Hamza, and M. B. Crawford, "Beemon: An iot-based beehive monitoring system," *Computers and Electronics in Agriculture*, vol. 190, p. 106427, Nov. 2021.

[40] W. Hong, B. Xu, X. Chi, X. Cui, Y. Yan, and T. Li, "Long-term and extensive monitoring for bee colonies based on internet of things," *IEEE Internet Things J.*, vol. 7, no. 8, 2020.

[41] F. Bellino, G. Turvani, U. Garlando, and F. Riente, "An integrated multisensor system for remote bee health monitoring," in *2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*. Perugia, Italy: IEEE, Nov. 2022, pp. 334–338.

[42] S. Aydin and M. N. Aydin, "Design and implementation of a smart beehive and its monitoring system using microservices in the context of iot and open data," *Comput. Electron. Agric.*, vol. 196, p. 106897, May 2022.

[43] S. Kontogiannis, "Beehive smart detector device for the detection of critical conditions that utilize edge device computations and deep learning inferences," *Sensors*, vol. 24, no. 16, p. 5444, Aug. 2024.

[44] U. Libal and P. Biernacki, "Drone flight detection at an entrance to a beehive based on audio signals," *Arch. Acoust.*, Jul. 2024.

[45] M. A. et al., "Urban: Urban beehive acoustics and phenotyping dataset," Jun. 2024.

[46] I. Nolasco, A. Terenzi, S. Cecchi, S. Orcioni, H. L. Bear, and E. Benetos, "Audio-based identification of beehive states," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 8256–8260.

[47] U. Libal and P. Biernacki, "Non-intrusive system for honeybee recognition based on audio signals and maximum likelihood classification by autoencoder," *Sensors*, vol. 24, no. 16, p. 5389, Aug. 2024.

[48] A. B. et al., "Bee together: Joining bee audio datasets for hive extrapolation in ai-based monitoring," *Sensors*, vol. 24, no. 18, p. 6067, Sep. 2024.

[49] A. Robles-Guerrero, T. Saucedo-Anaya, C. A. Guerrero-Mendez, S. Gómez-Jiménez, and D. J. Navarro-Solís, "Comparative study of machine learning models for bee colony acoustic pattern classification on low computational resources," *Sensors*, vol. 23, no. 1, p. 460, Jan. 2023.

[50] T. Zhang, S. Zmyslony, S. Nozdrenkov, M. Smith, and B. Hopkins, "Semi-supervised audio representation learning for modeling beehive strengths," May 2021.

[51] J. Yoo, R. Siddiqua, X. Liu, K. A. Ahmed, and M. Z. Hossain, "Beenet: An end-to-end deep network for bee surveillance," *Procedia Computer Science*, vol. 222, pp. 415–424, 2023.

[52] S. Jafor Sadeek Quaderi, S. Afrin Labonno, S. Mostafa, and S. Akhter, "Identify the beehive sound using deep learning," *arXiv e-prints*, pp. arXiv–2209, 2022.

[53] A. Zgank, "Iot-based bee swarm activity acoustic classification using deep neural networks," *Sensors*, vol. 21, no. 3, p. 676, 2021.

[54] S. Cecchi, A. Terenzi, S. Orcioni, P. Riolo, and N. Isidoro, *A Preliminary Study of Sounds Emitted by Honey Bees in a Beehive*, 2018.

[55] D. Fourer and A. Orlowska, "Detection and identification of beehive piping audio signals," in *7th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2022)*, 2022.

[56] A. Orlowska and D. Fourer, "Identification of beehive piping audio signals," 2021. [Online]. Available: https://dx.doi.org/10.21227/53mq-g936

[57] J. A. Calvo, "Open source beehives project - audio database as of 2-21-17," Feb 2017. [Online]. Available: https://doi.org/10.5281/zenodo.321345

[58] G. Duttakiit, "Smart bee colony monitor: Clips of beehive sounds (version 3)," 2023, accessed: Dec. 24, 2024. [Online]. Available: https://www.kaggle.com/code/gauravduttakiit/class-dataset-smart-bee-colony-monitor/notebook

[59] A. P. et al., "Pytorch: An imperative style, high-performance deep learning library," 2019.

[60] K. R. Bagadi and C. M. R. Sivappagari, "A robust feature selection method based on meta-heuristic optimization for speech emotion recognition," *Evolutionary Intelligence*, vol. 17, no. 2, pp. 993–1004, 2024.

[61] K. R. Bagadi, "A comprehensive analysis of multimodal speech emotion recognition," in *Journal of Physics: Conference Series*, vol. 1917, no. 1. IOP Publishing, 2021, p. 012009.

[62] I. Nolasco, A. Terenzi, S. Cecchi, S. Orcioni, H. L. Bear, and E. Benetos, "Audio-based identification of beehive states," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019, pp. 8256–8260.

[63] D. Braga, A. Madureira, F. Scotti, V. Piuri, and A. Abraham, "An intelligent monitoring system for assessing bee hive health," *IEEE Access*, vol. 9, pp. 89 009–89 019, 2021.

[64] H. Hadjur, D. Ammar, and L. Lefèvre, "Toward an intelligent and efficient beehive: A survey of precision beekeeping systems and services," *Computers and Electronics in Agriculture*, vol. 192, p. 106604, January 2022.

[65] A. Qandour, I. Ahmad, D. Habibi, and M. Leppard, "Remote beehive monitoring using acoustic signals," *Acoustics Australia*.

[66] D. Goyal and B. S. Pabla, "Condition based maintenance of machine tools—a review," *CIRP J. Manuf. Sci. Technol.*, vol. 10, pp. 24–35, Aug. 2015.

[67] A. V. et al., "Attention is all you need," Aug. 2023.

[68] M. Vaishnav, "Exploring the role of (self-)attention in cognitive and computer vision architecture," 2024.

[69] GitHub, "Librosa: Python package for music and audio analysis," https://github.com/librosa/librosa, Apr. 2024, accessed: Jan. 02, 2025.

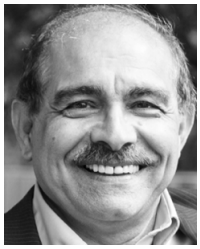
[70] K. R. Bagadi and C. M. R. Sivappagari, "An evolutionary optimization method for selecting features for speech emotion recognition," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 21, no. 1, pp. 159–167, 2023.

[71] U. Nations, "Sustainable development goals," https://sdgs.un.org/goals, 2015, accessed: Dec. 24, 2024.

SABA MUSTAFA received the B.S. and M.S. degrees in Computer Sciences from Lahore College For Women University, Lahore, Pakistan. She is currently pursuing a Ph.D. degree in Computer Science at Auckland University of Technology, Auckland, New Zealand. Her research focuses on the intersection of artificial intelligence, machine learning, and bioacoustics to study bee behaviour. She has over five years of experience in web development and has worked in higher education institutions in Pakistan and the Kingdom of Bahrain. Her research interests include smart environments for apiculture, signal processing, and the application of intelligent technologies to honey bee science.

MAHSA MOHAGHEGH (Senior Member, IEEE) was born in Iran. She received the bachelor's degree in computer engineering, the master's degree in computer architecture, and the Ph.D. degree in computer engineering from Massey University, in 2013. She is a computer engineer specializing in artificial intelligence and natural language processing in New Zealand. She is a Senior Lecturer and the Director of Women in Tech with the School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology. She has been involved with Google's Computer Science for High Schools Program since 2013 and runs workshops in Auckland. She founded She Sharp, a women's networking group aimed at encouraging girls and young women to engage with digital industries. She has received several awards, including the Emerging Leader category in the 2013 Westpac Women of Influence Awards, the 2019 YWCA Equal Pay Champion Award, and the 2020 Massey University Distinguished Alumni Award. She is a well-recognized leader in AI and machine learning and is committed to promoting diversity and inclusion in the tech industry.

IMAN T. ARDEKANI received the Ph.D. degree from the University of Auckland, Auckland, New Zealand. He is a Lecturer with the Computing and Information Technology Department, Unitec Institute of Technology, Auckland. He is the Founder of the Unitec Intelligent Rooms Laboratories, Auckland, where he is involved in signal and information processing algorithms. He has been granted several research awards and has published more than 40 papers in different journals and international conferences. His current research interests include adaptive control, statistical signal processing, and acoustics.



HOSSEIN A. SARRAFZADEH Abdolhossein Sarrafzadeh is currently a University Distinguished Professor and the Director of the Center of Excellence in Cybersecurity Research, Education, and Outreach (CREO), North Carolina Agricultural and Technical State University. He has worked in the areas of IoT, data mining, and machine learning in cybersecurity, with a focus on smart cities and smart grid security. He developed one of the world's first real-time facial expression and gesture recognition systems for emotion recognition. He has founded several cybersecurity research and operations centers and holds multiple patents in computer vision, cybersecurity, and cloud security. He has published more than 190 research articles and secured more than $30 M in funding in the past five years