Synergistic Mel-frequency Cepstral Coefficients and Short-time Fourier Transform for enhanced bee state detection using machine learning

Thi-Thu-Hong Phan[1]*

Department of Artificial Intelligence, FPT University, Danang, Vietnam

Abstract. This study investigates the potential of sound analysis to detect bee states within beehives, a critical challenge for beekeepers. We propose a novel approach combining Mel-Frequency Cepstral Coefficients (MFCC) and Short-Time Fourier Transform (STFT), effectively creating more informative representative data for the classification of bee states. Experimentation on a real dataset demonstrates that the Random Forest classifier utilizing synergy features extracted from MFCC and STFT outperforms models relying solely on MFCCs or STFT features, significantly improving classification accuracy (up to 87.2%). This integrated approach offers advantages in capturing both spectral detail (MFCC) and temporal information (STFT) potentially leading to improved classification accuracy for bee states detection. The findings contribute valuable insights for developing robust bee colony health monitoring systems.

Keywords: Bee sound · NoQueen · Bee state · STFT · MFCC · Machine learning methods· Combined features.

1 Introduction

Honey bees are essential pollinators, playing a critical role in maintaining ecosystem balance and enhancing agricultural yield. Their adept foraging supports the reproduction of plants, including key crops, and their hive products, such as honey and beeswax, cater to various human requirements. Furthermore, honey bee populations serve as indicators of overall environmental health, emphasizing their importance and the necessity for conservation measures.

Maintaining healthy bee colonies is crucial for beekeepers. They face constant threats like swarming, Varroa mites, and critically, missing queens. Traditionally, beekeepers rely on regular hive inspections for monitoring their hives. This involves physically opening the hive to visually inspect the colony's health, confirm the queen's presence, and assess her overall well-being, etc. However, these inspections are time-consuming, potentially disruptive to bee behavior and harmful to the queen. This requires a non-invasive approach to monitoring beehives and facilitating interventions. To address these limitations, many studies have been exploring advanced technology-based methods [3, 8]. The Internet of

* Corresponding author: hongptt11@fe.edu.vn

Things (IoT) and artificial intelligence play a crucial role in this new approach. Sensors-equipped beehives can continuously collect data on various environmental factors like temperature and humidity, and most importantly, these systems can capture bee sounds for further analysis.

Sound analysis, combined with machine learning algorithms, holds immense promise for beehive monitoring. Research has shown that this approach is effective in detecting bee states in bee colonies, with various algorithms being employed using sound as input data. For instance, Ruvinga et al. (2021) [13] achieved 92% accuracy using LSTM networks with Mel-frequency cepstral coefficients (MFCCs) to detect the queenless in the beehive. Similarly, Nolasco et al. (2019) explored both Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) with MFCCs, achieving high accuracy (up to 94% with the addition of Hilbert Huang Transform (HHT)). In [9], Phan et al. investigated new MFCC and hyper-parameters tuning techniques for enhanced performance of bee sound recognition. Truong et al. [14] developed an approach based on deep learning models to identify bee buzzing sounds from other sounds like noise or cricket chirping. Barbisan and Riente [1] achieved high accuracy (up to 98.8%) using both Neural Networks (NN) and Support Vector Machines (SVM) with 20 MFCC features. Further studies explored alternative methods for queen presence detection, Fourer et al. employed STFT with Convolutional Neural Networks (CNNs), achieving 96% accuracy. Similarly, Ho et al. (2023) utilized MFCC for feature extraction on a real dataset, reaching a peak accuracy of 91.75%.

In [12], Rustam et al. focused on classifying bee sounds into three categories: Bee, NoBee, and NoQueen. To achieve this goal, the authors employed a variety of feature selection techniques and machine learning algorithms. They achieved slightly higher accuracy with K-Nearest Neighbors (KNN) at 0.83 compared to 0.82 with Random Forest (RF). However, the performance of these models is still suboptimal on their dataset. This raises questions about their inefficiency and potential ways to improve their effectiveness. This challenge motivates us to find answers. Specifically, we propose a novel approach that combines Mel-Frequency Cepstral Coefficients (MFCC) and Short-Time Fourier Transform (STFT), effectively creating more informative representative data for detecting bee states in beehives. Experimental results show that these techniques can significantly improve the accuracy of machine learning methods in recognizing bee states within beehives.

The rest of the paper is structured as follows. Section 2 provides a concise description of the methods employed in the study. Section 3 presents the experiments conducted, the obtained results, and relevant discussions. Section 4 offers concluding remarks summarizing the key findings and potential future directions.

2 Methodology

This paper proposes a methodology for classifying various beehive states through audio analysis as figure 1. We leverage a combination of STFT and MFCC for

feature extraction to enhance the performance of ML methods for this detection task.

The process commences with gathering audio samples from beehives. These samples encompass a diverse range of bee activities, aiming to capture audio signatures associated with three distinct states: Bee presence, No bee presence, and No queen presence.

The collected audio samples are then processed to extract features suitable for bee state classification. This work employs two distinct techniques: STFT and MFCC. Each method offers valuable insights into the audio data – STFT revealing the time-frequency distribution and MFCC capturing the spectral envelope relevant to human hearing. Following extraction, the STFT and MFCC features are combined, creating a comprehensive representation of the audio data.

These combined features are subsequently fed into machine learning models for bee state classification. To ensure a comprehensive exploration of the data and identify the most suitable approach, we investigate six different machine learning algorithms: RF, ET (Extra Trees), XGBoost (eXtreme Gradient Boosting), SVM, KNN, and LR. Each algorithm brings its strengths allowing for a robust analysis of the beehive audio data. This exploration holds the potential to achieve the most accurate bee state classification results.

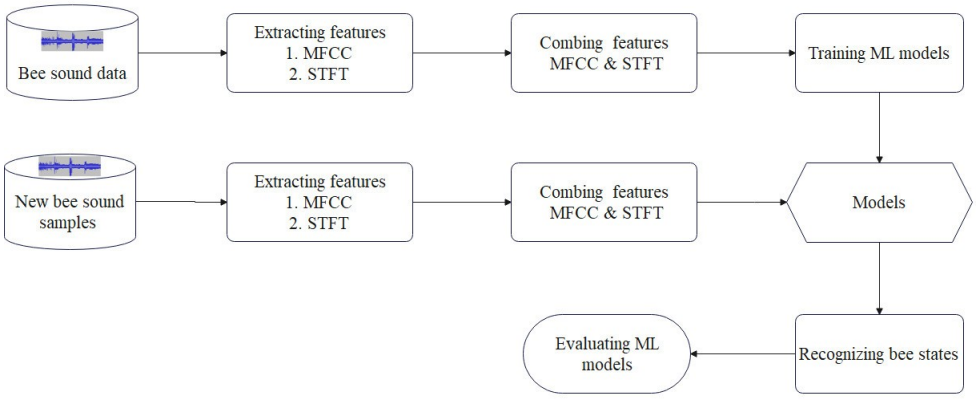

Fig. 1. Overview of proposed approach for recognizing bee states in beehives

2.1 Feature extraction methods

a) Short-Time Fourier Transform (STFT)

STFT is a fundamental method in signal processing, utilized across fields like audio analysis, speech recognition, and biomedical engineering [6]. It offers a time-frequency representation, enabling the examination of changing spectral

characteristics. Unlike standard Fourier analysis, which is suited for stationary signals, STFT performs localized Fourier transforms over small, overlapping windows, making it effective for dynamic signals. Figure 2 describes the steps involved in calculating STFT:

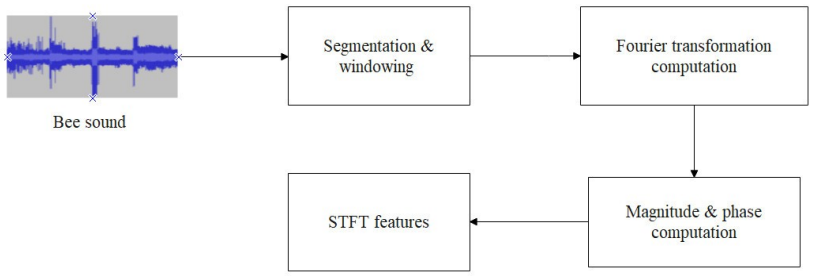

Fig. 2. Schema of STFT method

Divide the signal into short time segments: The signal is divided into short time segments, called frames, with a fixed length. The frame length determines the resolution of the STFT in time. A shorter frame length will provide better time resolution but lower frequency resolution, while a longer frame length will provide better frequency resolution but lower time resolution.

Apply a window function to each segment: A window function is applied to each segment to reduce artifacts caused by the abrupt truncation of the signal. Common window functions include Hannning, Hamming, and Gaussian windows. The choice of window function can affect the shape of the STFT peaks.

Compute the Fourier transform of each segment: The Fourier transform is computed for each segment to obtain its frequency spectrum. The Fourier transform decomposes the signal into its constituent sinusoidal components, revealing the signal's frequency content within that segment.

Combine the results: Each segment's magnitude and phase of the Fourier transform are combined to form the STFT. The magnitude represents the strength of each frequency component, while the phase represents the time delay of each frequency component.

Display the STFT: The STFT can be visualized as a time-frequency spectrogram, representing the magnitude by color intensity. The spectrogram shows how the frequency content of the signal changes over time.

b) Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a powerful method for extracting features from audio signals. This technique involves dividing the frequency band into sub-bands on the MEL scale and then applying the Discrete Cosine Transform (DCT) to extract Cepstral

Coefficients. The key steps of MFCC, as shown in Fig. 3, include pre-emphasis, framing, windowing, applying the Discrete/Fast Fourier Transform (DFT/FFT), Mel-frequency warping, and cepstrum calculation (inverse DCT).

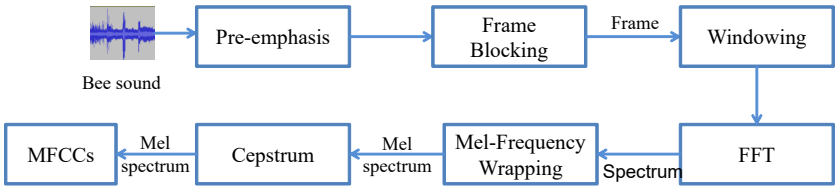

Fig. 3. Schema of MFCC algorithm

Pre-emphasis: The initial stage of MFCC, pre-emphasis, amplifies energy in higher frequencies [10]. This involves passing the signal through a first-order high-pass filter to reduce noise during sound capture.

Frame blocking and windowing: MFCC operates on short, stationary intervals of audio data by dividing the signal into overlapping frames. Each frame contains multiple audio samples with some overlap between consecutive frames. Hanning or Hamming windows are commonly applied [11] to enhance harmonics, smooth edges, and reduce edge effects during DFT/FFT computation.

Fast Fourier Transform (FFT) applying: In this step, each windowed frame is transformed into a magnitude spectrum using FFT, which quickly computes the Discrete Fourier Transform (DFT). This process converts each frame from the time domain to the frequency domain.

Mel-Frequency wrapping: The Mel spectrum is obtained by applying a Mel-filter bank to the power spectrum (derived from the signal's FFT). These Mel filters mimic human hearing, capturing the energy within specific frequency bands relevant to our perception of sound.

Cepstrum: Mel-scale power spectrum is then converted to the time domain using Discrete Cosine Transform (DCT) to obtain Mel-Frequency Cepstral Coefficients (MFCC).

In this study, we employed three different feature extraction parameters to extract MFCC features from the sound samples: 20 features align with the approach adopted by [1], and 40 and 80 features correspond to the feature extraction method used by [9].

c) Combining MFCC and STFT features

This study proposes a novel approach to enhance feature representation for bee state classification in beehives. We leverage the complementary strengths of two feature extraction techniques: Mel-Frequency Cepstral Coefficients and Short-Time Fourier Transform. MFCC excels at capturing the perceptual characteristics of sound, mimicking human hearing. This makes it particularly effective

for representing signals with prominent pitch content, like bee vocalizations. Additionally, MFCC offers dimensionality reduction, reducing computational complexity. On the other hand, STFT provides a detailed time-frequency representation of the signal. This allows it to capture subtle variations in frequency and transient events that might be crucial for distinguishing bee states in beehives. This characteristic makes STFT well-suited for analyzing non-stationary signals like beehive sounds, which often exhibit complex frequency patterns.

By strategically combining these features, we aim to create a more informative representation of the beehive audio data. This enriched representation is structured as:

$$stft_1, stft_2, \ldots, stft_{80}, mfcc_1, mfcc_2, \ldots mfcc_{40}$$

The new features incorporate both the perceptual and temporal aspects of the sounds, potentially leading to improved classification accuracy in identifying different bee states within colonies.

2.2 Machine learning models

K-Nearest Neighbors (KNN) is a popular supervised learning algorithm used for both classification and regression tasks [5]. It assigns a label to a new sample based on the labels of its K closest neighbors in the training set. The class label is typically determined by a majority vote among these neighbors, with closer neighbors potentially having more influence. KNN uses a distance metric, usually Euclidean distance, to measure proximity between data points. This method can easily adapt to new data without retraining the model. However, KNN performs poorly in high-dimensional spaces and is sensitive to noise and missing data in the training set.

Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification and regression tasks [7]. It aims to find the optimal hyperplane in a high-dimensional space that separates different classes with the maximum margin. This hyperplane is determined by support vectors, which are the critical data points that maximize class separation. SVM can handle linearly inseparable data using the kernel trick, mapping input features into higher-dimensional spaces for non-linear classification. In this study, we use the radial basis function (RBF) kernel exclusively, due to its demonstrated effectiveness in various scholarly investigations and publications.

Logistic Regression (LR)is a widely used method primarily for classification tasks. It works best when the relationship between the independent variables and the dependent variable (often binary) can be modeled linearly. LR is particularly useful for estimating the probability of an event occurring. The coefficients produced by LR indicate the extent to which each independent variable influences the likelihood of a specific outcome. Its simplicity and effectiveness make it a popular choice for binary classification problems where understanding variable influences is essential.

Random Forest (RF) is a powerful machine learning algorithm that combines multiple decision trees to improve accuracy and prevent overfitting [2].

Each tree in the forest is created using a random subset of the training data (bootstrapping), promoting diversity among the trees. The trees independently classify or predict outcomes by finding optimal splitting points, often using Gini impurity as a metric. For regression tasks, the predictions of all trees are averaged, while for classification tasks, the final prediction is determined by majority vote.

Extra Trees (ET) or Extremely Randomized Trees, is an ensemble learning algorithm similar to Random Forest (RF) but with some key differences. ET uses the entire original sample, which helps reduce bias. Additionally, ET introduces randomness by selecting split points randomly rather than computing the local optimum using metrics like Gini impurity or entropy. This random selection of split points increases diversity and reduces correlation among the trees, enhancing the algorithm's effectiveness. ET is known for its speed and ability to mitigate overfitting, making it well-suited for large datasets with many features.

XGBoost (XGB) or Extreme Gradient Boosting, is a highly efficient implementation of gradient-boosted decision trees designed for speed and performance [4]. It uses parallel boosting trees to smooth training loss and apply regularization, combining the strengths of multiple base trees for optimal results. The algorithm corrects previous mistakes, learning iteratively to improve performance. To reduce overfitting and accelerate training, XGBoost incorporates randomization techniques, such as selecting subsamples for tree construction and choosing features at various levels. It also employs percentiles to test a subset of candidate splits, significantly speeding up the process while maintaining accuracy, and making it effective for various data science problems.

3 Experiments

3.1 Data description

To evaluate the proposed approach, we use the dataset in the previous study [12]. This dataset, which comprises 13,792 samples, offers valuable insights into beehive activity through sound recordings, categorized into three classes:

Bee: Sounds generated by normal bee activity within the hive.

NoBee: represents ambient sound, indicating moments when external noise is present. When the queen is old, diseased, or deceased, workers may begin rearing new queens within 12-24 hours, leading to a decline in colony activity and sound levels. In such cases, only background or ambient noise is detected, suggesting potential colony collapse or rearing of a new queen.

NoQueen: Sounds associated with queenless hives, potentially signifying colony collapse (due to worker inactivity), queen rearing (leading to decreased activity), or other anomalies.

We divide this dataset into training and testing sets with an 80:20 ratio (table 1) to conduct and assess experiments.

Table 1. Sample distribution for Bee, NoBee, and NoQueen categories

Class	# Sample	Train (80%)	Test (20%)
Bee	5473	4378	1095
NoBee	3458	2766	692
NoQueen	4861	3889	972

3.2 Results and discussion

The performance of ML methods using individual features

a) MFCC features

Table 2 presents the performance of different machine learning models in identifying bee states (Bee, NoBee, and NoQueen) using MFCC with varying extraction features: 20 features, 40 features, and 80 features. The models considered include KNN, SVM, LR, RF, ET, and XGB. The table demonstrates the impact of the number of MFCC features on the performance of the machine learning models. In general, a model trained with more relevant features tends to achieve higher accuracy, but adding too many or irrelevant features can lead to overfitting and reduced performance. This suggests that extracting more detailed spectral information from the sound recordings can enhance the ability to distinguish between different bee states. RF indicates the most substantial improvement in accuracy with an increasing number of features, reaching a peak of 83.74% with 40 features. This suggests that RF is highly sensitive to feature selection and may benefit from optimization between 20 and 40 features. ET exhibits a similar trend to RF, with a peak accuracy of 84.05% at 40 features and a slight decrease with 80 features.

In [12], Rustam et al. employed a combination of feature selection techniques (PCA, Chi-squared, and SVD) applied to a large set of 1740 MFCC features, achieving the highest accuracy of 83% for bee detection using RF and KNN classifiers. In our study, we utilize a simpler approach that direct extraction of a smaller set of MFCC features (40 features) can achieve a comparable or even better accuracy with 84.05%.

Table 2. Performance of ML methods for identifying bee states using three different MFCC feature sets (%)

Model	20 features	40 features	80 features
KNN	79.46	81.49	82.41
SVM	66.05	68.63	69.53
LR	69.45	75.33	75.64
RF	81.73	83.74	83.54
ET	82.21	84.05	83.18
XGB	80.96	82.58	83.33

b) STFT features

Table 3 presents the test accuracy achieved by different machine learning models in identifying bee states using Short-Time Fourier Transform (STFT) features extracted with two different feature sets: 80 features and 257 features. In general, all models except SVM show a slight increase in accuracy with more features (from 80 to 257 features). This suggests that capturing a wider range of frequency information can enhance the ability to distinguish between different bee states. KNN exhibits the most significant improvement (1.83%), followed by RF (0.56%) and ET (0.31%). LR also shows a noticeable improvement (2.88%). Consistent with the previous analysis using MFCC features, XGBoost again achieves the highest accuracy across both feature counts (85.40% with 80 features and 85.72% with 257 features). This reinforces its robustness and effectiveness in bee state identification.

When comparing the performance of ensemble ML models (RF, ET, and XGB) using STFT features and MFCC features, it is evident that using STFT features, even with 80 features, they outperform the best performance achieved using MFCC features (84.05%). This indicates that STFT features are more effective in capturing the relevant information for bee state identification compared to MFCC features in this specific case.

Table 3. Performance of ML methods for identifying bee states using two STFT features set (%)

Model	**80 features**	**257 features**
KNN	81.66	83.49
SVM	75.01	75.91
LR	68.41	71.29
RF	84.75	85.31
ET	84.10	84.41
XGB	85.40	85.72

The performance of ML methods using combined features

The previous analyses have demonstrated the effectiveness of both MFCC and STFT features in bee state identification. MFCC features excel in capturing the perceptual characteristics of sound, while STFT features provide a detailed representation of the frequency spectrum. To harness the strengths of both MFCC and STFT features, we propose a feature fusion approach that combines the extracted features from both methods. We opt to combine MFCCs with 80 STFT features after evaluating the trade-off between performance and computational efficiency. While using 257 STFT features yielded slightly higher accuracy, the significant increase in features also led to higher computational

costs. To achieve a balance between performance and efficiency, we chose 80 STFT features.

Table 4 presents the accuracy achieved by different machine learning models in identifying bee states using a combination of 80 STFT features and two different sets of MFCC features: 20 features and 40 features. All models show an improvement in accuracy when using 40 MFCC features compared to 20 MFCC features in combination with 80 STFT features. This suggests that incorporating more perceptual frequency content of sound captured by additional MFCC features can enhance the performance of most models. RF and ET demonstrate the most significant improvement in accuracy with 40 MFCC features, with increases of 1.18% and 0.53% respectively. XGBoost exhibits a minimal improvement in accuracy (0.27%) with 40 MFCC features compared to 20 features. This suggests that XGBoost might already be effectively utilizing the information provided by both STFT and 20 MFCC features. KNN and SVM exhibit a moderate increase in accuracy with 40 MFCC features, suggesting they can benefit from additional information but to a lesser extent compared to models like RF and ET. Logistic Regression (LR) shows a noticeable improvement in accuracy (3.84%) with 40 MFCC features, suggesting it utilizes the additional temporal information effectively.

Figure 4 provides a clear and concise illustration of how feature fusion leverages the strengths of different feature extraction techniques like MFCC and STFT. By combining the perceptual information captured by MFCC features with the frequency information captured by STFT features, the model gains a richer understanding of the audio data, leading to more accurate bee state classification. The combination of MFCC (either 20 or 40 features) with STFT features (80) reaches the highest point on the chart. This visually emphasizes that combining features captures a more comprehensive representation of the beehive sounds, leading to improved model performance in identifying bee states.

In the previous study [12], the highest accuracy achieved was around 83%. By employing feature fusion, the new approach has pushed the accuracy to 87.20%, representing a substantial leap forward. This improvement suggests that the combination of MFCC and STFT features captures more relevant information about beehive sounds, enabling the model to better distinguish between different bee states.

Table 4. Performance of ML methods for identifying bee states using 80 STFT with different MFCC feature sets (%)

Model	**80 STFT & 20 MFCC features**	**80 STFT & 40 MFCC features**
KNN	80.14	82.75
SVM	65.97	68.85
LR	75.28	79.12
RF	86.01	87.19
ET	86.44	86.97
XGB	86.68	86.95

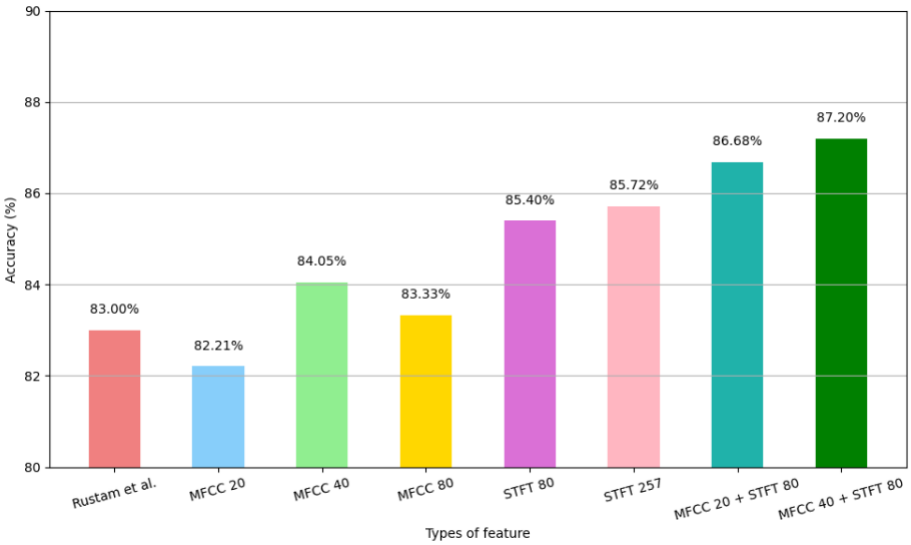

Fig. 4. Comparison of the best accuracy using different feature types

4 Conclusion

This paper proposes a novel approach for improved bee state recognition by generating more informative representative data through the fusion of MFCC and STFT features. We extract and analyze both MFCC and STFT features individually, and subsequently combine them to create new representative data that encompasses a richer set of information about beehive sounds. Experiment results show that RF model achieved a remarkable accuracy of 87.2%, surpassing the previous best result by a significant margin of 4.2% [12]. These findings strongly support the effectiveness of the proposed approach in generating representative data that significantly enhances bee state recognition performance. While this result represents a significant improvement, it still falls short of the ultimate goal of achieving highly accurate and reliable bee state recognition. Recognizing the need for further advancements, we propose investigating the application of state-of-the-art deep learning techniques, such as transfer learning and RegNet modes, to address this challenge.

References

1. Barbisan, L., Riente, F.: Machine learning framework for the acoustic detection of the queen bee presence. Acta Acustica (2023)
2. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)

3. Cecchi, S., Spinsante, S., Terenzi, A., Orcioni, S.: A Smart Sensor-Based Measurement System for Advanced Bee Hive Monitoring. Sensors **20**(9), 2726 (May 2020)
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory (1967)
6. Durak Ata, L., Arikan, O.: Short-time fourier transform: Two fundamental properties and an optimal implementation. Signal Processing, IEEE Transactions on **51**, 1231 – 1242 (06 2003). https://doi.org/10.1109/TSP.2003.810293
7. Evgeniou, T., Pontil, M.: Support vector machines: Theory and applications. Machine Learning and Its Applications, Advanced Lectures (2001)
8. Liao, Y., McGuirk, A., Biggs, B., Chaudhuri, A., Langlois, A., Deters, V.: Non-invasive Beehive Monitoring through Acoustic Data Using SAS® Event Stream Processing and SAS® Viya®. SAS Global Forum p. 24 (2020)
9. Phan, T.T.H., Nguyen-Doan, D., Nguyen-Huu, D., Nguyen-Van, H., Pham-Hong, T.: Investigation on new mel frequency cepstral coefficients features and hyper-parameters tuning technique for bee sound recognition. Soft Computing (2022). https://doi.org/10.1007/s00500-022-07596-6
10. Picone, J.: Signal modeling techniques in speech recognition. Proceedings of the IEEE **81**(9), 1215–1247 (Sep 1993)
11. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (Feb 1989). https://doi.org/10.1109/5.18626, conference Name: Proceedings of the IEEE
12. Rustam, F., Zahid Sharif, M., Aljedaani, W., Lee, E., Ashraf, I.: Bee detection in bee hives using selective features from acoustic data. Multimedia Tools and Applications **82**(5), 7095–7112 (2023). https://doi.org/10.1007/s11042-023-15192-5
13. Ruvinga, S., Hunter, G.J., Duran, O., Nebel, J.C.: Use of LSTM Networks to Identify "Queenlessness" in Honeybee Hives from Audio Signals. In: 2021 17th International Conference on Intelligent Environments (IE). pp. 1–4. IEEE (Jun 2021). https://doi.org/10.1109/IE51775.2021.9486575
14. Truong, T.H., Nguyen, H.D., Mai, T.Q.A., Nguyen, H.L., Dang, T.N.M., Phan, T.T.H.: A deep learning-based approach for bee sound identification. Ecological Informatics **78**, 102274 (2023). https://doi.org/10.1016/j.ecoinf.2023.102274