Bioacoustic Dataset of Sudden Queen Loss in an *Apis mellifera L.* Honeybee Colony

Ícaro de Lima Rodrigues[1], **Davyd B. de Melo**[1], **Danielo G. Gomes**[1]

[1]Programa de Pós-Graduação em Engenharia de Teleinformática
Grupo de Redes de Computadores Engenharia de Software e Sistemas (GREat)
Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza - CE

icarodelima@alu.ufc.br, {davydmelo, danielo}@ufc.br

Abstract. *The queen bee plays a crucial role in maintaining colony well-being. However, traditionally determining her presence within a hive requires invasive methods, causing stress among the bees and potential nest damage. Audio monitoring offers a simpler, cost-effective, and non-invasive approach to detect various conditions using colony bioacoustics. In this study, we recorded an Africanized honeybee colony for six days, removing the queen on the final day. From these recordings, we sampled 300 frames of 3 seconds each, extracting three time-domain features, five spectral features, and 13 Mel-Frequency Cepstral Coefficients (MFCCs). This resulted in a dataset of 5,400 samples capable of distinguishing queen presence or absence, even with class imbalance.*

1. Summary table

Keywords: Honeybees, bioacoustics, feature extraction, dataset;

WCAMA 2024 topic: Precision beekeeping, biodiversity;

Data type: Tabular – 23 columns and 5,400 rows;

Brief data description: Twenty-one audio features were extracted from recordings collected during a six-day period, simulating a sudden queen loss event in a honeybee colony. The recordings were captured using a basic smartphone earphone.

Data format: Comma separated values (.csv) file;

Collection site: Apiary of the Bee Sector in Department of Zootechnics at Universidade Federal do Ceará (UFC) - approximately 3.7425°S, 38.5789°W.

Public repository: Kaggle and Mendeley Data.

2. Objectives

The primary goal of the proposed dataset is to support the development of machine learning algorithms and pattern recognition techniques beneficial for beekeeping. These tools aim to extract relevant insights from bioacoustics, aiding in the identification of queen absence in healthy hives and enabling swift resolutions. Furthermore, microphones, being simple and cost-effective equipment, advocate for audio monitoring as an efficient and non-invasive solution. As specific objectives, the dataset intends to (i) explore techniques for handling unbalanced data in classification problems, such as data augmentation and resampling; (ii) achieve accurate queen presence classification using anomaly detection, one-class classification, or incremental learning methods; and (iii) be openly accessible in a data repository for public use.

3. Material and Methods

We conducted audio recordings using a healthy colony of Africanized honeybees in an observation hive at UFC Bee Sector, located on Pici Campus in Fortaleza-CE. Figure 1 illustrates an overview the observation hive and experimental setup.

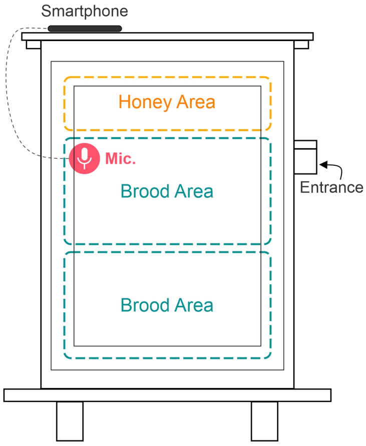

Figure 1. Observation hive showing colony areas and microphone placement.

Recordings were made consistently over six days, from 1 PM to 5 PM each day, with a 44.1 kHz frequency sampling rate and WAV format. Initially, the queen was present for the first five days but was removed from the hive early in the morning of the sixth day. The presence of the queen and the overall health status of the colony were confirmed with the assistance of a local beekeeper, who also removed the queen. As a result, audio samples were obtained, allowing for a distinction between the positive class (y^+: queen present) and negative class (y^-: queen absent). Meteorological variables for the data acquisition period were obtained from the Meteostat[1] website, as outlined in Table 1.

Table 1. Meteorological variables during the period of recording.

		Meteorological variables^{*2}			
Date	Queen	T (ºC)	RH (%)	AP (hPa)	Ws (km/h)
2019-08-26	present	29.3	56.8	1013	26.3
2019-08-27	present	29.4	56.0	1013	30.0
2019-08-28	present	29.6	55.8	1012	26.3
2019-08-29	present	30.1	48.2	1012	27.1
2019-08-30	present	28.9	57.8	1013	29.7
2019-09-02	absent	29.1	64.4	1011	29.3

* The full names of the variables were abbreviated. **T** stands for temperature, **RH** for relative humidity, **AP** for air pressure, and **Ws** for wind speed. These data were obtained in Meteostat website.

To efficiently process the extensive audio recordings, each of the six files was divided into 300 random 3-second segments, thereby reducing the original four-hour files

[1] https://meteostat.net/

to 900 seconds each. Subsequently, feature extraction was performed using the pyAudioAnalysis [Giannakopoulos 2015] library in Python. The extracted features included three time-domain features, five spectral features and 13 Mel-Frequency Cepstral Coefficients (MFCCs), selected based on their common use in machine learning studies on bee sounds [Abdollahi et al. 2022], as well as their availability in the library.

Initially, features were calculated for short-term windows, with windows of 50 ms size and a 25 ms overlap. Afterward, these short-term features were converted into mid-term (segment-level) features. For the mid-term computation, we selected non-overlapping 1-second segments, with each second of recording corresponding to one sample in the dataset. Thus, 900 seconds per day of recording were utilized to extract one sample per second into the dataset, resulting in a total of 5,400 samples. Figure 2 illustrates a parallel coordinates plot for visualizing the multidimensional data distribution. Additionally, Principal Component Analysis (PCA) was applied to reducing dimensionality and create effective visualizations, such as the scatter plot depicted in Figure 3.

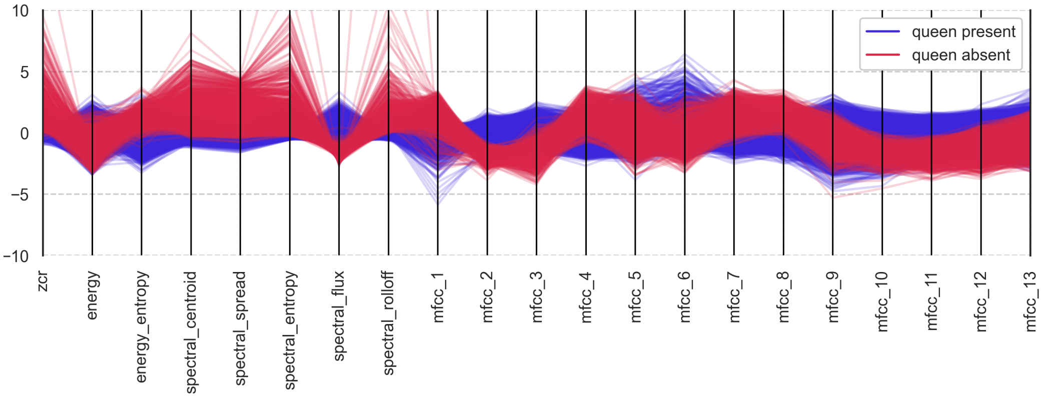

Figure 2. Parallel coordinates visualization. Blue lines represent queen present samples, while red lines represent queen absent samples.

4. Data availability

The .csv file containing the 21 audio features, along with the date of recording and label, and comprising 5,400 samples from the six raw audio files, is available in a public repository [Lima Rodrigues et al. 2024b]. The total file size is 2.15 MB.

5. Conclusion

Monitoring colonies via bioacoustic patterns offers a cost-effective alternative without compromising efficiency. Our dataset supports this, showing satisfactory results at a lower financial cost for beekeepers. Using the One-Class SVM classifier [Pimentel et al. 2014] we achieved 96% accuracy in our preliminary assessment, as depicted in [Lima Rodrigues et al. 2024a]. Incremental learning also show promise, with 97% accuracy [Rodrigues et al. 2022]. Although honeybee species are extensively studied in precision beekeeping, much of the research has focused on species in the United States and United Kingdom [Abdollahi et al. 2022]. While these studies provide valuable insights into bee behavior and colony health, there is a need for further research encompassing a broader range of bee species and geographical locations, such as Africanized honeybees in tropical climates. Finally, to our knowledge, this is the first public dataset based on the acoustic patterns produced by Africanized honeybees.

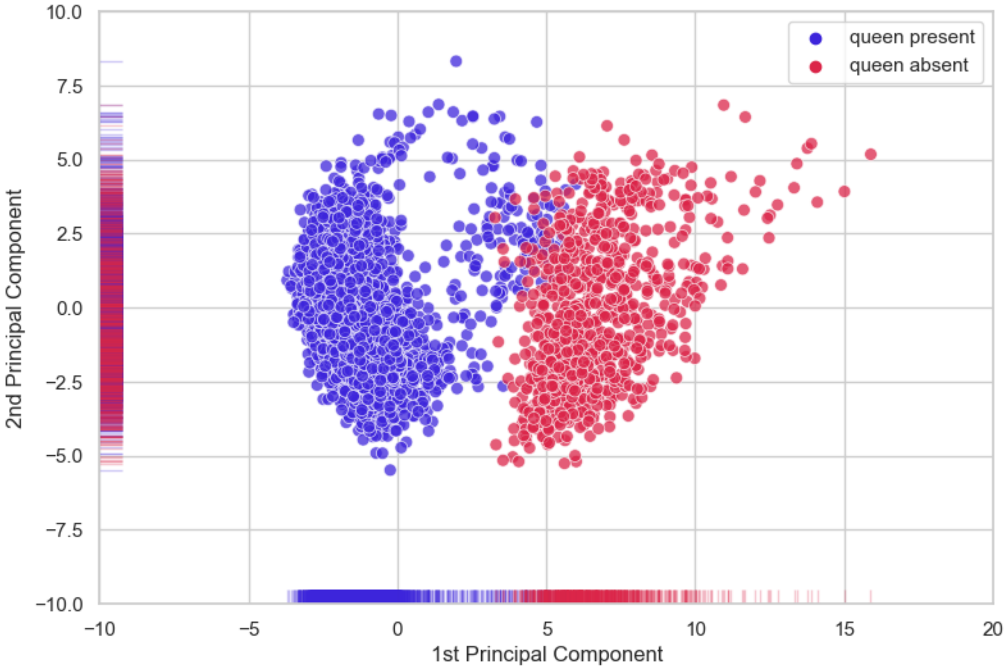

Figure 3. Scatterplot (markers) and marginal distribution (colored axis ticks) of the dataset represented by the 2 Principal Components after performing PCA.

Acknowledgements

This research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001. Danielo G. Gomes thanks the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant number #311845/2022-3). Ícaro de Lima Rodrigues also thanks the support of the CNPq (process number: 140696/2023-7).

References

[Abdollahi et al. 2022] Abdollahi, M., Giovenazzo, P., and Falk, T. H. (2022). Automated beehive acoustics monitoring: A comprehensive review of the literature and recommendations for future work. *Applied Sciences*, 12(8).

[Giannakopoulos 2015] Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PLOS ONE*, 10(12):1–17.

[Lima Rodrigues et al. 2024a] Lima Rodrigues, Í., Melo, D. B., and Gomes, D. G. (2024a). Queen loss event in africanized honeybee colony. [link].

[Lima Rodrigues et al. 2024b] Lima Rodrigues, Í., Melo, D. B., and Gomes, D. G. (2024b). Sudden queen loss event in an africanized honeybee colony. *Mendeley Data*. [link].

[Pimentel et al. 2014] Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal processing*, 99:215–249.

[Rodrigues et al. 2022] Rodrigues, Í. L., Melo, D., Silva, D., Rybarczyk, Y., and Gomes, D. (2022). Padrões bioacústicos como identificadores precisos da presença de rainha em colmeias de abelhas melíferas. In *Anais do XIII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 11–20. SBC.