scientific data

Check for updates

OPEN

DATA DESCRIPTOR

UrBAN: Urban Beehive Acoustics and PheNotyping Dataset

Mahsa Abdollahi[1], Yi Zhu[1], Heitor R. Guimarães[1], Nico Coallier[2], Ségolène Maucourt[3], Pierre Giovenazzo[3] & Tiago H. Falk[1]✉

In this paper, we present a multimodal dataset obtained from a honey bee colony in Montréal, Quebec, Canada, spanning the years of 2021 to 2022. This apiary comprised 10 beehives, with microphones recording more than 3000 hours of high quality raw audio, and also sensors capturing temperature, and humidity. Periodic hive inspections involved monitoring colony honey bee population changes, assessing queen-related conditions, and documenting overall hive health. Additionally, health metrics, such as Varroa mite infestation rates and winter mortality assessments were recorded, offering valuable insights into factors affecting hive health status and resilience. In this study, we first outline the data collection process, sensor data description, and dataset structure. Furthermore, we demonstrate a practical application of this dataset by extracting various features from the raw audio to predict colony population using the number of frames of bees as a proxy.

Background & Summary

Honey bees (*Apis mellifera*) play a critical role in ecological balance, serving as essential pollinators for both agricultural crops and natural biodiversity. Their contributions extend beyond honey and beeswax production, impacting many fruit and seed crops, including almonds, citrus fruits, and blueberries. This dependence underscores the significance of honey bee populations for food production sustainability and quality. However, declines in bee colony health and population size can have far-reaching consequences for the agricultural industry.

Conventional methods of beehive surveillance rely on manual and visual inspections, which are labor-intensive for beekeepers and disruptive to colonies, resulting in infrequent checks. Typically, beekeepers examine their hives every two weeks during active periods of pollination or honey production, and less frequently during winter. However, significant changes in colony dynamics and health can occur within this time frame, thus necessitating continuous monitoring. Global reports of substantial colony losses underscore the urgency of this issue.

In recent years, there has been a global observation of significant colony losses, which have been attributed to various stressors working either independently or in combination. These stressors include pesticides, pathogens, parasites, climate variations, as well as other factors[1–5]. Consequently, passive monitoring of honeybee colony health has attracted significant attention from the beekeeping and the research communities.

Recently, with the advancement of IoT (Internet-of-Things) in precision beekeeping, automated beehive monitoring tools have emerged to overcome the shortcomings of human manual inspections and colonies management[6]. These systems typically deploy sensors within the hive to monitor real-time colony status and assess its condition[7–9]. Existing systems typically gather information such as temperature, humidity, beehive weight, and acoustics. For example, temperature stability within the hive is crucial for bee health and brood development, thus directly impacting hive productivity[10–12]. Moreover, relative humidity affects larval growth, colony development, and bee behavior, with variations influencing water transportation and feeding[13,14]. Moreover, hive weight is an essential measurement for researchers, offering insights into colony activities, such as nectar collection and food consumption, showing variations over the course of the day. Continuous monitoring of colony weight, particularly a reference colony, aids in identifying the start and end of nectar flow and in assessing colony foraging activity[15,16].

While temperature and humidity can provide some complimentary information about the health of a colony, beehive acoustics have proven to be a more effective method, as bees communicate internally using vibrations and acoustic signals, generated through body movements, wing flapping, and muscle contractions[17]. These signals

[1]INRS-EMT, Université du Québec, Montréal, Canada. [2]Nectar Technologies Inc., Montréal, Canada. [3]Département de biologie, Université Laval, Quebec City, Canada. ✉e-mail: tiago.falk@inrs.ca

Name	# Hives	Labels	Modality	Size	Availability
NU-Hive[33]	2	Queenright/queenless	Raw audio	96 h	Public
			Temperature	—	Not available
			Humidity	—	Not available
BUZZ[32]	6	Bee buzzing/cricket/noise	Raw audio	7 h	Upon requests
OSBH[23]	6	Queenright/queenless	Raw audio	140 min	Subset available
Too bee or not to bee[23]	8	Bee buzzing/no bee buzzing	Raw audio	12 h	Public
MSPB[34]	53	Population; Honey yield; Queen conditions; Hygienic behavior; Winter mortality	Hand crafted audio features	Quarter-hourly during 365 day	Public
			Temperature		
			Humidity		
Smart Bee Colony Monitor[52]	4	Queenright/queenless	Raw audio	118 h	Public
			Temperature	Hourly during 40 day	
			Humidity		
			Pressure		
UrBAN (ours)	10	Population; Queen conditions; Winter mortality	Raw audio	3171 h	Public
			Temperature	Quarter-hourly during 135 day	
			Humidity		

Table 1. Comparison of UrBAN dataset with the other public datasets.

include sounds associated with different events, such as mite attacks, queen failure, and swarming, making them an ideal modality for beehive monitoring. In the literature, various acoustic monitoring systems have been highlighted, offering capabilities such as queen absence detection[18–21], bee activity[22,23], swarming[24–26], hive strength[27], pathogen/parasite infestations[28], environmental pollutants[29,30], and early prediction of colony winter survivability[31].

Open access data on beehive management is crucial for fostering research output and advancements in the field, especially with the recent advances seen in artificial intelligence (AI) and machine learning (ML). Unfortunately, the beekeeping community faces a significant scarcity of open access data, hindering the progress of research and innovation. Access to comprehensive datasets on beehive dynamics, including factors such as colony health, behavior, and environmental influences, empowers researchers to conduct in-depth analyses and develop impactful solutions. While there are some public bee audio database available such as Buzz[32], Nu-hive[33], and OSBH[23], they are mostly focused on bee and queen bee detection and are limited in sample size. To bee or not to bee dataset[23] is a combination of Nu-hive and osbh audio samples, labeled for a different purpose which is detecting bee buzzing sound.

Recently, the MSPB dataset was released[34]; this is a longitudinal multi-sensor dataset with phenotypic measurements has been released showing that how audio signals can be used for detection of winter survivability and population estimation. The dataset, however, does not provide access to the raw audio signals and makes available only pre-processed parameters extracted from the audio signal (e.g., overall hive audio power). Table 1 lists available beehive acoustic datasets along with some relevant statistics. It can be seen from Table 1 that the existing datasets are limited in size of raw audio samples and the days that they covered. Therefore, to address these limitations and support advancements in beehive monitoring, we present a new beehive acoustics dataset.

In this paper, we introduce the UrBAN (Urban Beehive Acoustics and PheNotyping) dataset that includes over 3000 hours of raw audio samples collected from beehives during a period of two years. The main focus of the data was on colony population prediction. The population of a colony can be estimated using the number of frames of bees covered by least 70%[35]. There have been some studies showing that the number of frames of bees can be predicted using various features extracted from the raw audio within a machine learning framework[34,36,37]. Authors in[34] showed that features such as audio power in specific band of frequency and its variation can be used in predicting hive population.

The UrBAN dataset was gathered over the period spanning 2021 to 2022, originating from observations made across a network of ten beehives, part of an urban apiary, positioned in the rooftop of a building located in Montreal, Canada. Various parameters such as audio recordings, temperature measurements, and relative humidity readings were collected. A subset of the audio recordings has been previously analyzed in[36], where the focus was on predicting the number of frames of bees and investigating the impact of environmental noise. In contrast, this study provides a comprehensive description of the dataset, highlighting its structure, features, and potential applications. Notably, the dataset encompasses a broad spectrum of critical metrics pertaining to hive inspections, including assessments of colony honey bee population dynamics, evaluations of queen-related conditions, and records of overall health status. Specific health indicators such as *Varroa* mite infestation rates and assessments of winter survivability are also cataloged, providing invaluable insights into the factors influencing hive health and resilience.

Methods

Urban Apiary. The rooftop apiary situated in Montréal, Canada (45.5253°N, 73.6123°W), comprised ten active honeybee hives. These hives were placed on wooden pallets (2 hives per pallet) in a row facing south east. Figure 1a shows the apiary placement. Each hive consisted of one brood chamber and one to two honey supers, housed within 10-frame standard Langstroth boxes, with a maximum of three boxes per hive. All hives originated from four-frame nucs, which were acquired and installed during the month of May in the year 2021.

(a) (b)

Fig. 1 Photos of the (**a**) rooftop urban apiary in Montreal and (**b**) audio recording hardware.

Year	Inspections	Varroa mite measurment	Audio recording	Temperature/Humidity
2021	June 22th–October 19th	—	August 11th–October 31st	June 19th – October 31st
2022	July 11th–September 7th	August 24th, September 1st, and September 30th	February 1st–October 31st	—

Table 2. Summary of the data collection start and end dates for each year.

Experienced beekeepers deliberately maintained varied hive populations to capture diverse data points, aligning with the primary objective of predicting beehive strength, which correlates with population size. At the beginning of the data collection, each hive contained a distinct number of (full) frames of bees with a minimum of six frames of bees in hives with a single brood box to a maximum of 20 frames in hives with both a brood box and a honey super. As colony populations expanded over time, additional honey supers were introduced. Consequently, within our apiary, the maximum configuration comprised three boxes and up to 30 frames of bees.

Hive Management and Inspection. The hives were manually inspected roughly every two weeks to measure the strength of the hives (i.e., the number of frames of bees), to verify the presence of a laying queen, as well as to report any additional observations related to the colony activity. The start and date of these inspections are listed in Table 2 for each year. Figure 2 shows the histogram of the number of frames of bees observed during multiple inspections for each year of the experiments. Each histogram indicates the count of observed frames of bees for all of the beehives during the experiments. Moreover, Fig. 3 shows the number of frames of bees for each hive during inspections. The beehive colonies were wintered outdoors with the aid of insulation, a strategic approach aimed at enhancing their survivability and well-being during the colder months, as shown in Fig. 4. Insulation provides an additional layer of protection against harsh environmental conditions, helping to maintain stable temperatures within the hive and reduce heat loss.

Varroa Mite Infestation. Varroa mite infestation poses a significant threat to the health and well-being of honeybee colonies, making it a matter of utmost importance for beekeepers and researchers alike. In the 2022 data collection, the beekeepers measured the varroa mite infestation using alcohol wash method[38] in some of the beehives. The alcohol wash method is a common technique for measuring Varroa mite infestations in honeybee colonies. It involves collecting a sample of about 300 bees from the brood nest, submerging them in isopropyl alcohol (70% or higher), and gently shaking the container to dislodge the mites. The mixture is then strained, and the mites are counted to estimate the infestation rate. This method, while resulting in the loss of the sampled bees, provides an accurate assessment of mite levels, helping beekeepers make informed decisions about mite control and ensuring colony health. Figure 5 indicates the amount of varroa mite infestation in each measurement.

Mortality Rate. Winter mortality is one of the primary causes of beehive losses globally and is a significant concern for both beekeepers and researchers[39]. As temperatures drop and resources become scarce, honeybee colonies face numerous challenges that can impact their survival. Factors such as disease prevalence, mite infestations, inadequate food stores, and harsh environmental conditions all contribute to increased mortality rates among bee colonies during the winter[40]. Understanding and monitoring these mortality rates is crucial for assessing the health of bee populations and implementing strategies to mitigate losses. During the data collection, 20% of the beehives died after overwintering in 2022.

Sensor Data. A multimodal sensor, positioned atop the central frame within the bottom brood box, facilitated continuous monitoring of internal hive temperature and humidity (SHTC1, Beecon, Nectar Technologies Inc, Canada (https://www.nectar.buzz/)). This sensor operates within a relative humidity range of 0% to 100% RH, with

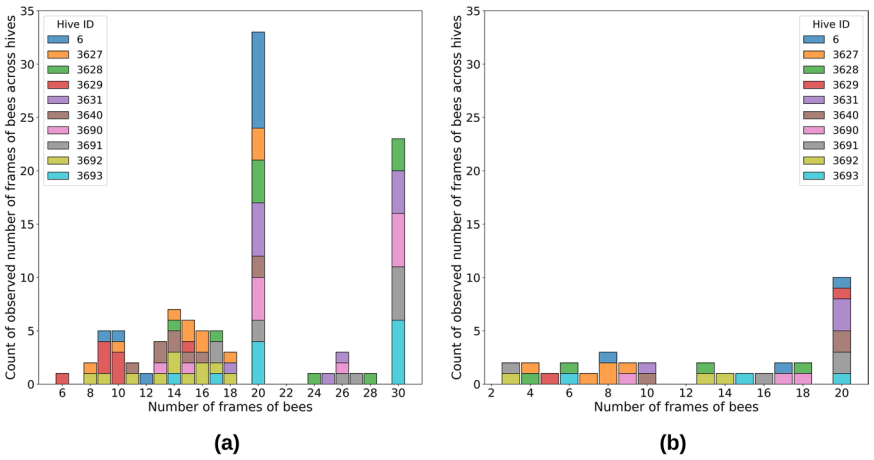

Fig. 2 Histograms of the observed number of frames of bees for the year of (**a**) 2021 and (**b**) 2022 experiments. Each hive is distinguished by a unique hive ID, which is presented in the figures' legend. The y-axis indicates the count of samples (inspections) for each number of frames of bees per hive.

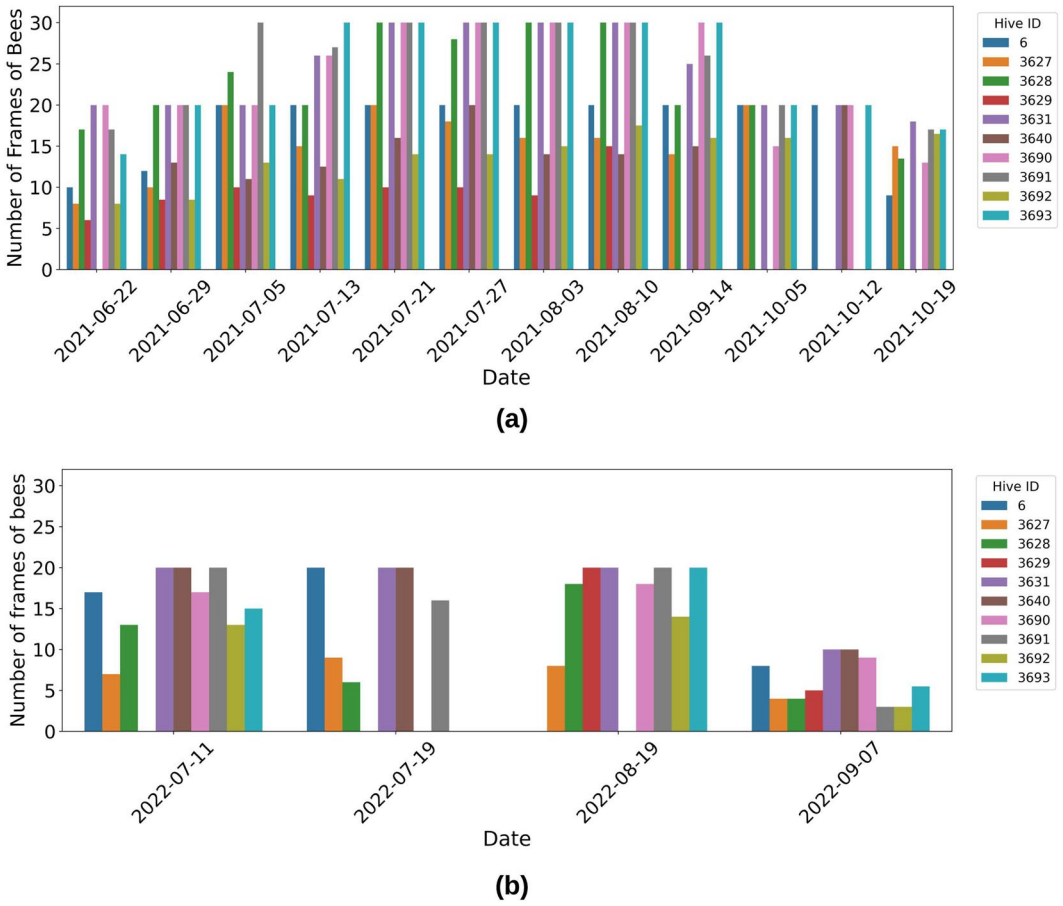

Fig. 3 Barplots of the number of frames of bees on days of inspections for the year of (**a**) 2021 and (**b**) 2022 experiments.

an accuracy of ±3% RH. Its temperature range spans from −40°C to +125°C, offering a measurement accuracy of ±0. 3°C. Additionally, an accompanying microphone, depicted in Fig. 6b, was installed adjacent to the sensor. The microphones were installed inside a custom-designed enclosure, which served to minimize direct contact with the bees and reduce the likelihood of propolization. The setup utilized a MEMS microphone (SPK0641HT4H-1) with a frequency range of 20 Hz to 20 kHz. This omnidirectional PDM microphone operates at a voltage range of 1.6 V to 3.6 V and has a sensitivity of −26 dB ± 1 dB @ 94 dB SPL. The multi-modal data is comprised of the

Fig. 4 Insulation used for beehives overwintering.

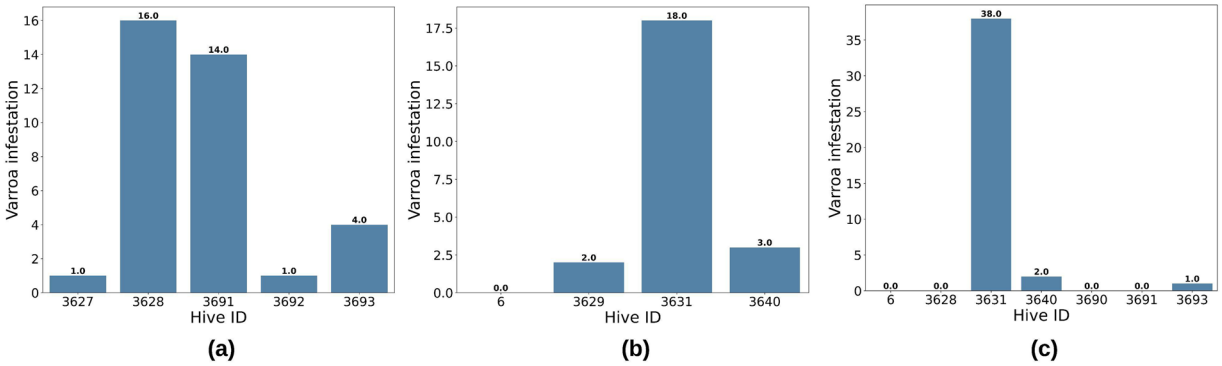

Fig. 5 Varroa mite infestation on (**a**) August 24th, (**b**) September 1st, and (**c**) September 30th, 2022.

Fig. 6 Location of the (**a**) temperature and humidity sensor placed on top of the middle frame of the first brood chamber and (**b**) microphone placed next to it.

average temperature and humidity readings every 15 minutes, and a 15-minute audio segment every 30 minutes with a sampling rate of 48 kHz. To minimize storage usage, every audio file was downsampled to 16 kHz.

Moreover, local external temperatures, humidity, and rainfall amounts levels were obtained from the Environment and Climate Change Canada website (https://climate.weather.gc.ca/). A representative example of a 24 h snapshot of the changes in internal/external temperature and humidity levels for a single beehive, as well as a 4-month period average of all beehives is shown in Fig. 7a,b, respectively. The 24 h snapshot (7a) is for a strong and healthy colony in August with one brood chamber and 2 honey supers with a total of 30 frames of bees (covered with at least 70% of bees).

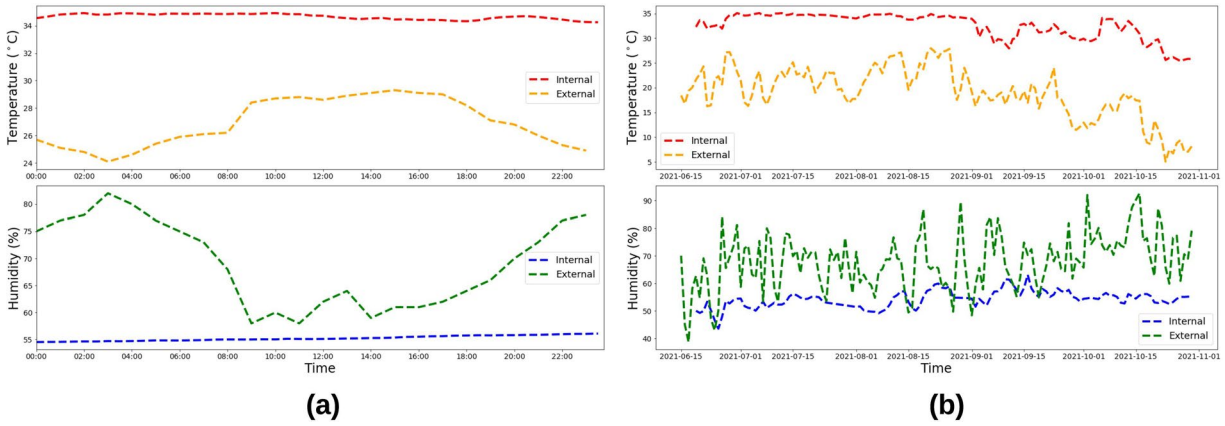

Fig. 7 Internal and external temperature and humidity during (**a**) a 24 h period for a single beehive on August 12th, 2021, and **b**) total duration of the experiment in 2021.

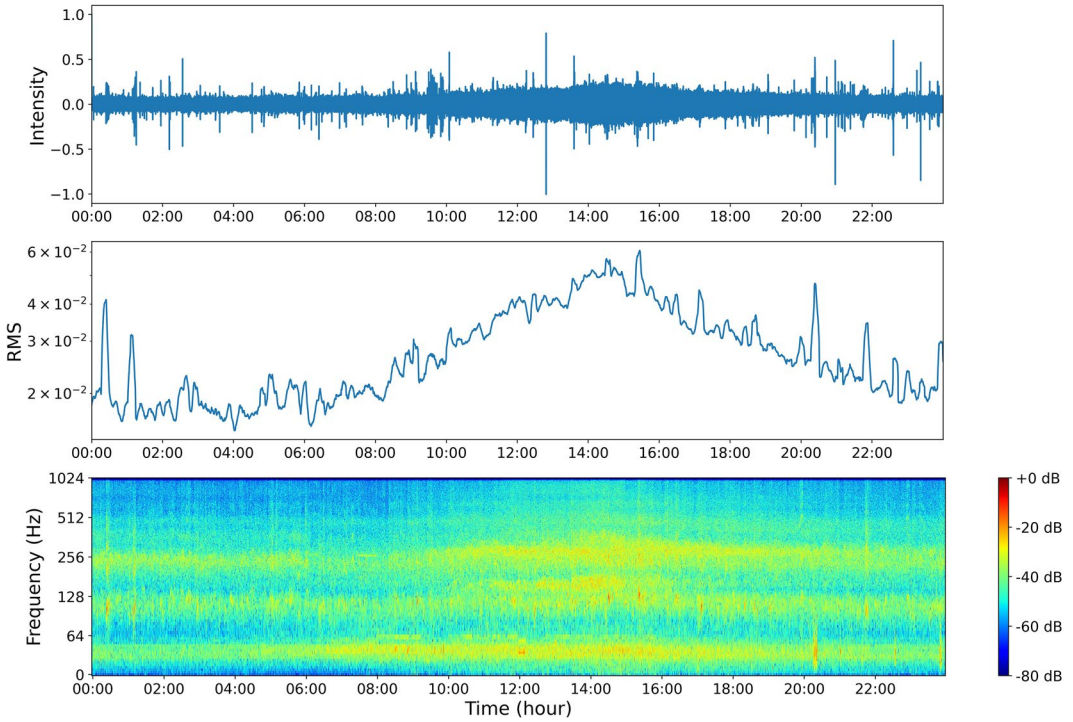

Fig. 8 Audio intensity, audio RMS, and corresponding spectrogram for a hive with with 3 full boxes of bees (August 16, 2021).

Figure 8 illustrates the intensity of the audio, the smoothed root mean square (RMS) power, and its corresponding spectrogram. It is evident from the figure that the audio power experiences a rise throughout the day, particularly during periods characterized by rising external temperatures and declining humidity levels. This observation suggests increased foraging activity and thermohygrometric regulation within the colony.

On the importance of beehive size and its effect on audio, Fig. 9 shows the 24 h bar-plots for different number of frames of bees. Each of these plots show the average RMS value with specified number of frames of bees. Similar to Fig. 8, a rising trend during the day can be seen. Table 3 provides an overview of the quantity and size of raw audio collected for each year. It consists of two columns detailing the total duration in hour and size in gigabytes (GB) of recordings, with one column encompassing all recordings regardless of inspection periods, and the other focusing solely on recordings made during the inspection periods specified in Table 2.

Data Records

The UrBAN dataset is made fully publicly available at the Federated Research Data Repository[41] (https://doi.org/10.20383/103.0972). Tables 4 and 5 provide a comprehensive overview of the inspections (labels) and sensor data, respectively. The inspection files called inspections_2021.csv and inspections_2022.csv contain details such as the count of frames of bees, presence of varroa mite infestation, queen status, and mortality rates stored

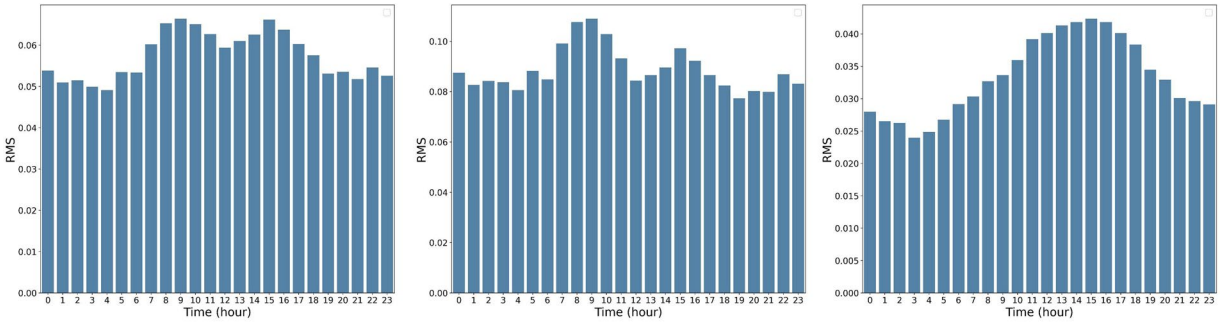

Fig. 9 24h average of RMS value when the number frames of bees is between (**a**) 10 and 20, (**b**) 20 and 30, and (**c**) 1 and 30.

Year	Type	Total quantity - size	Quantity during months of inspections
2021	Raw audio	1752 h - 185 Gb	1466 h - 155 Gb
2022	Raw audio	12491 h - 1315 Gb	1705 h - 180 Gb

Table 3. The quantity of raw audio for each year of the experiments.

as .csv files. Each beehive is identified by a unique hive ID, enabling discrimination between them. In Table 5, the details about recording of internal temperature, humidity, raw audio, and also weather information such as external temperature, humidity, and amount of precipitation are described. The raw audio recordings are wav files stored in a compressed format for easy download. The sensor data and weather information comprise two .csv files.

Technical Validation

Audio Enhancement. Removing environmental noise from bee acoustic audio is crucial for enhancing the effectiveness of monitoring systems in beekeeping. Environmental noise can obscure the sounds produced by bees, making it difficult to accurately detect and analyze important behaviors and events within the hive. Moreover, it could reduce the accuracy of the monitoring system significantly. While some studies explored methods in removing noises, such as human speech[42,43], there is still a need for a general noise removal step.

Spectral amplitude subtraction is a technique used in audio processing to enhance the quality of audio recordings. As shown in the block diagram in Fig. 10, it involves subtracting the spectral components of noise or unwanted signals from the original audio signal to reduce background noise and improve signal clarity. By identifying the spectral profile of noise or interference and subtracting it from the audio signal, spectral amplitude subtraction helps isolate the desired sound, resulting in cleaner and more intelligible audio recordings. Figure 10 shows the diagram of the spectral subtraction algorithm, where $y(n)$ and $\hat{x}(n)$ are the noisy and cleaned signals, respectively. As the figure indicates, after framing the audio and calculating the fast Fourier transform (FFT), the noise spectral profile needs to be estimated and eventually subtracted from the noisy signal, following the equation:

$$|\hat{X}_m(\omega)| = |Y_m(\omega)| - |D_m(\omega)|. \tag{1}$$

Here, we used an exponential moving average (EMA) filter to estimate the noise.

$$d[n] = (1 - \alpha)y[n] + \alpha d[n - 1], \tag{2}$$

where $d[n]$ and α are the noise signal and the weighting factor or smoothing parameter, respectively. In order to only update α in noisy frames, an adaptive algorithm is used[44]. Using this method, α will be a function of a-posteriori signal-to-noise ratio (denoted as $SNR_{\text{a-posteriori}}$) and is calculated for each frame, i.e.,

$$\alpha(m) = \frac{1}{1 + exp^{-a(SNR_{a-posteriori}(m)-T)}}, \tag{3}$$

where the parameter a represents the steepness of the sigmoid function's slope and T is a threshold value for the $SNR_{\text{a-posteriori}}$. The parameter T determines the point at which the smoothing parameter α transitions between its extremes in response to varying $SNR_{\text{a-posteriori}}$ values. A larger T causes the adaptive algorithm to identify a frame as noisy only if the $SNR_{\text{a-posteriori}}$ is considerably low, thereby reducing α. Conversely, a smaller T makes the algorithm more sensitive to detecting noise. The parameter T must be tuned based on the characteristics of the input signals and the noise environment.

Year	File name	Columns	Description
2021	inspections_2021.csv	Date	Time stamp (YYYY-MM-DD)
		Tag number	ID unique to each hive
		Colony size	The number of boxes for each beehive
		Fob 1st	The number of frames of bees in the first box
		Fob 2nd	The number of frames of bees in the second box
		Fob 3rd	The number of frames of bees in the third box
		FoBrood	The number of frames of brood
		Frames of honey	The number of frames of honey
		Queen status	QR/QNS (queen seen or not seen)
		Open	Time stamp indicating opening the box for inspections (HH:MM)
		Close	Time stamp indicating closing the box after inspections (HH:MM)
		Note	Additional observation such as beehive being weak or aggressive
2022	inspections_2022.csv	Date	Time stamp (YYYY-MM-DD HH:MM:SS)
		Tag number	ID unique to each hive
		Category	Hive grading, hive status, frames of bees, varroa, treatment, feeding, custom practice, queen management, hive issues
		Action detail	Detail of each category. Hive grading: 'strong', 'medium', 'weak', 'pulled honey super', 'size - 1d'; Hive status: 'queenright', 'queenless', 'deadout'; frames of bees: the number of frames of bees; Varroa: the varroa mite measurement; Treatment: 'mite away'; Feeding: 'sugar'; Custom practice: 'add entrance reducer', 'supering', 'added bee escape', 'added trash bag (feeder trick)'; Queen management: 'potential breeder'; Hive issues: 'chalk brood'
		Queen status	Queenright/queenless
		Is alive	0/1 (Zero indicates a dead hive)
		Report notes	Additional observation such as beehive being weak or aggressive

Table 4. Structure of the files describing inspections for each year.

The $\mathrm{SNR}_{\text{a-posteriori}}$ is estimated using the power spectrum of the noisy signal and the mean noise power spectrum over the past N frames, i.e.,:

$$\mathrm{SNR}_{\text{a-posteriori}}(m) = \frac{|Y_m(\omega)|^2}{\frac{1}{N}\sum_{k=m-N}^{m-1}|D_k(\omega)|^2}, \tag{4}$$

where $|Y_m(\omega)|^2$ is the power spectrum of the noisy signal in the current frame, and the denominator represents the average noise power spectrum estimate over the last N frames, effectively smoothing out noise variations. The estimation of the $\mathrm{SNR}_{\text{a-posteriori}}$ provides an indication of how much noise is present in each frame. The $\mathrm{SNR}_{\text{a-posteriori}}$ can be interpreted as follows:

- When $\mathrm{SNR}_{\text{a-posteriori}} \approx 1$, the frame predominantly contains noise.
- For $\mathrm{SNR}_{\text{a-posteriori}} > 1$, the frame contains both speech and noise.

The scripts for the spectral subtraction algorithm are made available at at our Github repository (https://github.com/mahsa-abdollahi/UrBAN) to facilitate study replication. Figure 11 depicts the audio signal collected from a hive before and after enhancement. The spectrograms illustrate distinct energy bands at approximately 64 Hz, 128 Hz, and 256 Hz both before and after spectral amplitude subtraction. In the noisy signal's spectrogram, noise is shown to overlap with frequency bands characteristic of bee audio. Following spectral amplitude subtraction, the processed spectrogram exhibits a reduction in noise while retaining the essential frequency components of the bee audio.

Feature Extraction. A machine learning framework for predicting hive strength through audio analysis comprises several essential stages, including signal measurement, pre-processing, feature extraction, and regression. Following the enhancement of the audio signal and the removal of unwanted noise, feature extraction becomes crucial. In this process, four distinct feature sets are derived for predicting the state of bee audio frames: mel-frequency cepstral coefficients (MFCCs), linear-frequency cepstral coefficients (LFCCs), spectral shape descriptors, and some hand-crafted parameters described in[34].

MFCCs have emerged as a cornerstone in audio-based applications, replicating the auditory processing mechanism of the human ear by employing mel-scale frequency mapping before cepstrum analysis[45]. Within the realm of precision beekeeping, these features have garnered considerable attention, prominently featured in approximately 30% of studies examined in previous reviews, particularly for tasks such as bee and queen presence, as well as swarming detection[7]. Here, 12 coefficients are extracted alongside the zeroth coefficient, utilized as a measure of signal power using 26 mel filters. Additionally, LFCCs are extracted through linear filters for comparison with MFCCs.

Spectral shape descriptors play a crucial role in the analysis and characterization of audio signals. In this paper, nine spectral shape descriptors are computed, including centroid, spread, skewness, kurtosis, entropy,

Year	File/folder name	Columns	Description
2021	sensor_2021.csv	Date	Time stamp (YYYY-MM-DD HH:MM:SS)
		Tag number	ID unique to each hive
		Temperature	Internal temperature in degree Celsius
		Humidity	Internal humidity in percentage
	audio_2021	—	Audio files names: "DD-MM-YYYY_HHhMM_HIVE_Tag.wav" where 'Tag' is the hive ID number.
2022	audio_2022	—	Audio files names: "DD-MM-YYYY_HHhMM_HIVE_Tag.wav" where 'Tag' is the hive ID number.
2021–2022	weather_2021_2022.csv	Date/Time (LST)	Time stamp (YYYY-MM-DD HH)
		Temp (°C)	External temperature in degree Celsius
		Rel Hum (%)	External humidity in percentage
		Wind Spd (km/h)	The speed of wind
		Precip. Amount (mm)	The amount of precipitation

Table 5. Structure of the files/folders of the sensor data (temperature and humidity), raw audio recordings, and weather information.

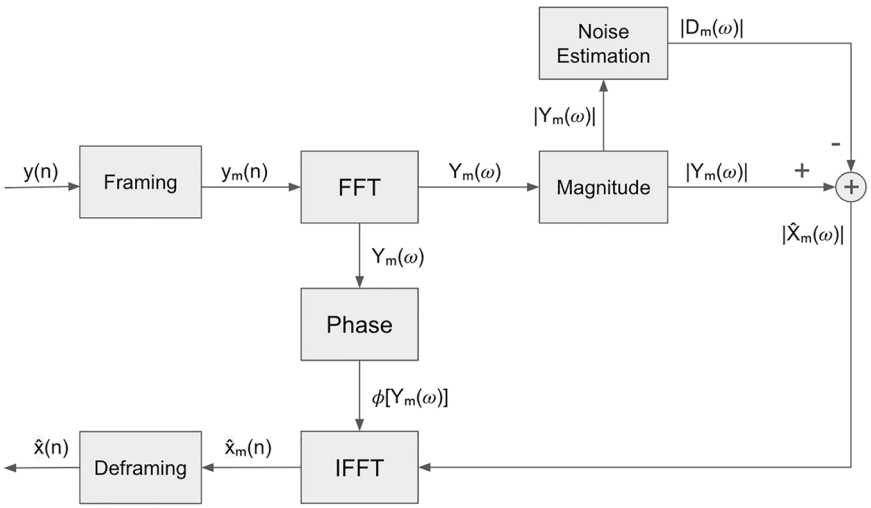

Fig. 10 Block diagram of the spectral amplitude subtraction.

rolloff, flatness, crest, and flux. Furthermore, other works have relied on hand-crafted audio features, including hive power (power between 122 Hz and 515 Hz), audio band density ratio (the ratio of hive power to the power of the entire frequency range), audio density variation (reflecting changes in hive power within each audio frame), and audio band coefficients in 16 linearly spaced frequency bins. For further details, interested readers are directed to[34]. Scripts to extract all the features tested herein are made available on our Github repository (https://github.com/mahsa-abdollahi/UrBAN) to facilitate experiment replication.

Validation: Number of Frame of Bees Prediction. Here, we use number of frames of bees prediction as a task to validate the dataset under an ML framework. In accordance with recommendations from previous studies[21,36], two distinct experimental configurations are explored: "random-split" and "hive-independent". The complete audio dataset spanning 18 (one of the beehives had a problem in audio recording) hives from the years 2021 and 2022 is used for the experiment. Initially, 15% of the dataset was set aside exclusively for feature selection to identify the most relevant features, ensuring that this process did not bias the model evaluation (or leak to the test set). The remaining 85% of the data was then used for model training and testing. For this, in the random-split approach, 5-fold cross-validation was used to ensure robust evaluation of the models. This cross-validation process ensures that every data point is included in the test set exactly once, reducing the risk of overfitting and providing a more reliable estimate of model performance. Conversely, in the hive-independent setup, 4 hives were set aside for feature selection and the 5 folds are selected hive-independently for training and testing using the remaining hives.

In the domain of ML, feature selection is important to avoid the curse of dimensionality and to improve the generalization abilities of models. By carefully selecting features, redundant or irrelevant attributes can be excluded, thereby reducing the risk of overfitting and improving the model's capacity to identify meaningful patterns within the data. To achieve this goal, several feature selection techniques were explored, including random forest feature importance[46,47], Principal Component Analysis (PCA)[48], minimum Redundancy Maximum

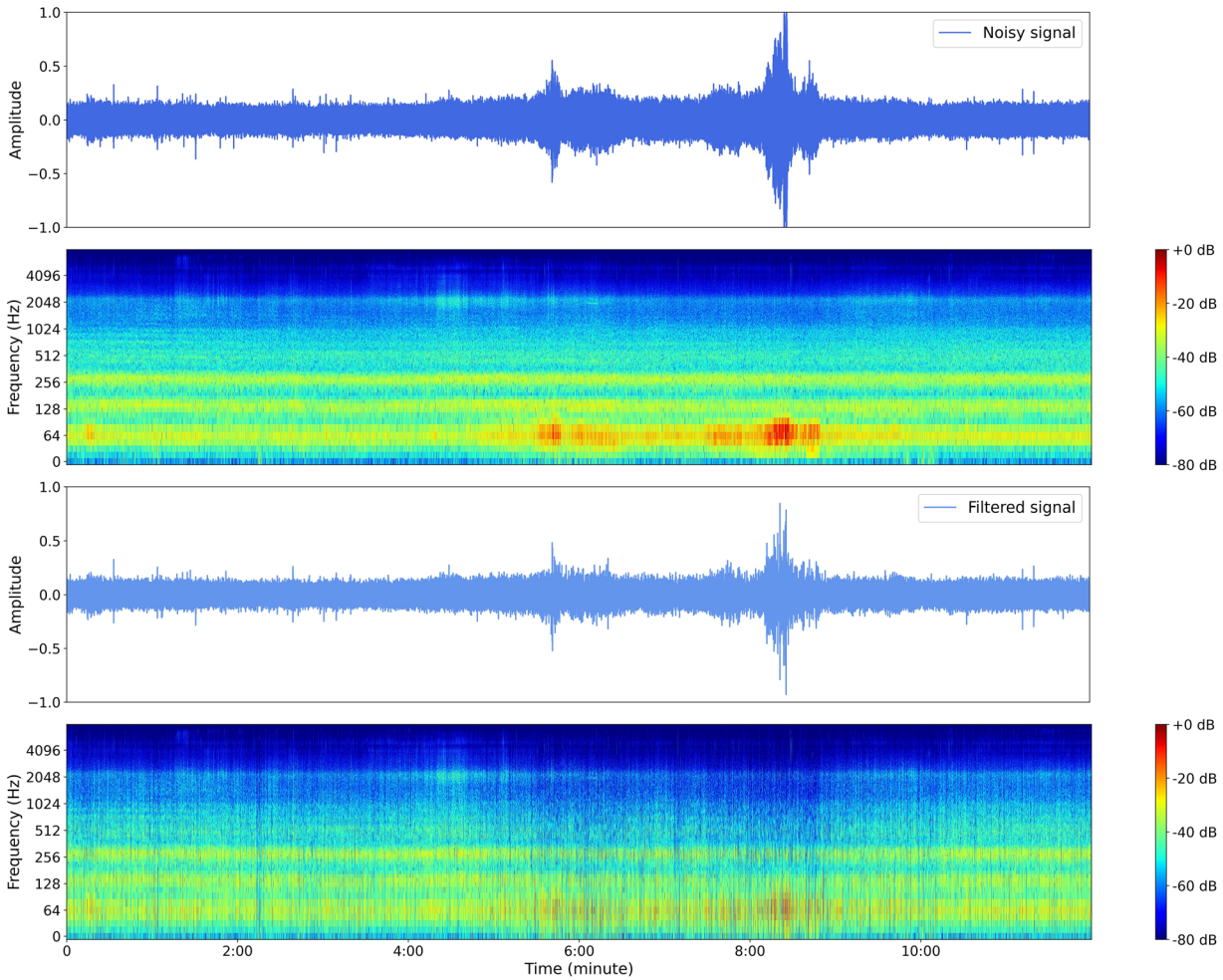

Fig. 11 From top to bottom, noisy audio amplitude of a beehive, the filtered audio, and their corresponding spectrograms.

Features	Random-Split					
	No pre-processing			Spectral amplitude subtraction		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Random baseline	3.63 ± 0.07	4.62 ± 0.08	21.47 ± 0.51	3.63 ± 0.07	4.62 ± 0.08	21.49 ± 0.43
LFCCs	0.65 ± 0.05***	1.63 ± 0.10***	4.78 ± 0.33***	0.60 ± 0.02***	1.53 ± 0.04***	4.44 ± 0.13***
MFCCs	0.41 ± 0.01***	1.21 ± 0.07***	3.10 ± 0.12***	0.42 ± 0.02***	1.18 ± 0.18***	3.12 ± 0.11***
Spectral descriptors	1.79 ± 0.06***	3.09 ± 0.08***	11.68 ± 0.44***	1.66 ± 0.06***	2.95 ± 0.08***	10.94 ± 0.45***
Hand-crafted	1.84 ± 0.05***	3.11 ± 0.07***	12.23 ± 0.33***	1.65 ± 0.04***	2.94 ± 0.03***	11.10 ± 0.53***
	Hive-Independent					
Random baseline	4.27 ± 1.86	5.17 ± 1.83	26.19 ± 4.41	4.23 ± 1.88	5.14 ± 1.83	26.15 ± 4.58
LFCCs	4.06 ± 1.51ns	4.91 ± 1.36ns	24.59 ± 5.01ns	4.01 ± 1.37ns	4.84 ± 1.25ns	23.63 ± 3.39ns
MFCCs	3.71 ± 0.58†	4.75 ± 0.49ns	25.65 ± 8.00ns	3.61 ± 0.68*	4.54 ± 0.63*	25.00 ± 8.70ns
Spectral descriptors	3.67 ± 1.09†	4.77 ± 1.01ns	23.81 ± 6.05†	3.55 ± 0.95**	4.44 ± 0.96*	22.30 ± 8.46*
Hand-crafted	4.16 ± 1.30ns	4.89 ± 1.17ns	25.64 ± 3.25ns	4.10 ± 1.35ns	4.89 ± 1.22ns	25.14 ± 6.48ns

Table 6. Performance comparison of error metrics between different feature sets and data partitioning setups, with and without spectral enhancement. The significance of p-values from the t-test is indicated as follows: $p < 0.001$ is indicated as *** (very significant), $p < 0.01$ is indicated as ** (significant), $p < 0.05$ is indicated as * (moderately significant), $p < 0.1$ is indicated as † (marginally significant), and $p \geq 0.1$ is indicated as ns (not significant).

Features	Random-Split			
	No pre-processing		Spectral amplitude subtraction	
	Correlation	R^2	Correlation	R^2
Random baseline	0.88 ± 0.00	0.53 ± 0.01	0.88 ± 0.00	0.54 ± 0.00
LFCCs	0.97 ± 0.00***	0.94 ± 0.00***	0.97 ± 0.00***	0.94 ± 0.00***
MFCCs	0.98 ± 0.00***	0.96 ± 0.00***	0.98 ± 0.00***	0.96 ± 0.00***
Spectral descriptors	0.89 ± 0.00ns	0.79 ± 0.00***	0.90 ± 0.00***	0.81 ± 0.00***
Hand-crafted	0.88 ± 0.00ns	0.78 ± 0.00***	0.90 ± 0.00***	0.81 ± 0.00***
	Hive-Independent			
Random baseline	0.52 ± 0.20	− 1.06 ± 3.16	0.52 ± 0.20	− 1.05 ± 3.16
LFCCs	0.60 ± 0.20ns	− 0.71 ± 1.79ns	0.62 ± 0.18ns	− 0.64 ± 1.66ns
MFCCs	0.66 ± 0.05*	0.15 ± 0.89*	0.68 ± 0.06**	0.20 ± 0.95*
Spectral descriptors	0.62 ± 0.14†	0.16 ± 0.41†	0.63 ± 0.12*	0.25 ± 0.40*
Hand-crafted	0.60 ± 0.15ns	− 0.98 ± 2.43ns	0.62 ± 0.13ns	− 0.80 ± 2.12ns

Table 7. Performance comparison of correlation and R^2 between different feature sets and data partitioning setups, with and without spectral enhancement. The significance of p-values from the t-test is indicated as follows: $p < 0.001$ is indicated as *** (very significant), $p < 0.01$ is indicated as ** (significant), $p < 0.05$ is indicated as * (moderately significant), $p < 0.1$ is indicated as † (marginally significant), and $p \geq 0.1$ is indicated as ns (not significant).

Relevance (mRMR)[49], and SHAP (SHapley Additive exPlanations[50]. Each feature set underwent testing with these methods, and the most effective approach was determined based on performance metrics evaluated on the validation set. Subsequently, a random forest regressor was employed to predict the number of frames of bees. Model evaluation was conducted using five key metrics: mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), Pearson correlation, and R-squared (R^2) of the predictions relative to the ground truth values.

Tables 6 and 7 presents the performance of each feature set both before and after pre-processing/enhancement under the two different split scenarios. To ensure the significance of the results, a random baseline regressor is also employed. In this random baseline regressor, MFCCs features are used, but the number of frames of bees are randomized between 0 and 30 during training. Each metric is presented as an average ± standard deviation with its corresponding p-value from the t-test, indicating the statistical significance of the model's performance compared to a random baseline. In the "random-split" case, the MFCCs outperform the baseline model and also other features both with or without pre-processing. In the "hive-independent" case, all features have lower performance compared to the "random-split", while the spectral descriptors achieve the best results. On the importance of audio enhancement and removing noise, the results indicated that in most of the experiments, the performance improved after spectral amplitude subtraction.

Usage Notes

The four CSV files including, inspections_2021, inspections_2022, sensor_2021, and weather_2021_2022 can be easily read using Python's Pandas library. An example code can be found in the scripts used for creating the plots in this paper and also the feature extraction and regression. In order to use the raw audio recordings which are stored as wav files, we recommend using the Python's Librosa library. An example of this procedure is found in the scripts related to feature extraction step available at our Github repository (https://github.com/mahsa-abdollahi/UrBAN).

This dataset provides different labels, as detailed in Table 4. This can enable the development of different supervised learning tasks. Moreover, given the advances seen with unsupervised learning, specifically self-supervised learning (SSL) of audio signals[51], the dataset can open door for numerous other applications. The work in[27,43], for example, used SSL techniques to detect beekeeper speech in the beehive audio recordings to improve hive monitoring performance. It is hoped that the UrBAN dataset will enable new applications that can improve the work of the beekeepers and the lives of the honey bees.

Code availability

The code used for creating the plots, audio enhancement, feature extraction, and number of frames of bees prediction are all categorized and available at our Github repository (https://github.com/mahsa-abdollahi/UrBAN).

Received: 4 July 2024; Accepted: 20 March 2025;
Published online: 31 March 2025

References

1. French, S. *et al*. Honey bee stressor networks are complex and dependent on crop and region. *Current Biology* **34**, 1893–1903 (2024).
2. Gaubert, J., Giovenazzo, P. & Derome, N. Individual and social defenses in Apis mellifera: a playground to fight against synergistic stressor interactions. *Frontiers In Physiology*. **14** (2023).
3. Brodschneider, R. *et al*. Multi-country loss rates of honey bee colonies during winter 2016/2017 from the COLOSS survey. *Journal Of Apicultural Research* **57**, 452–457 (2018).

4. Gray, A. *et al*. Loss rates of honey bee colonies during winter 2017/18 in 36 countries participating in the COLOSS survey, including effects of forage sources. *Journal Of Apicultural Research* **58**, 479–485 (2019).
5. Gray, A. *et al*. Honey bee colony winter loss rates for 35 countries participating in the COLOSS survey for winter 2018–2019, and the effects of a new queen on the risk of colony winter loss. *Journal Of Apicultural Research* **59**, 744–751 (2020).
6. Cota, D., Martins, J., Mamede, H. & Branco, F. BHiveSense: An integrated information system architecture for sustainable remote monitoring and management of apiaries based on IoT and microservices. *Journal Of Open Innovation: Technology, Market, And Complexity* **9**, 100110 (2023).
7. Abdollahi, M., Giovenazzo, P. & Falk, T. Automated beehive acoustics monitoring: A comprehensive review of the literature and recommendations for future work. *Applied Sciences* **12**, 3920 (2022).
8. Terenzi, A., Cecchi, S. & Spinsante, S. On the importance of the sound emitted by honey bee hives. *Veterinary Sciences* **7**, 168 (2020).
9. Alleri, M. *et al*. Recent developments on precision beekeeping: A systematic literature review. *Journal Of Agriculture And Food Research*. pp. 100726 (2023).
10. Cetin, U. The effects of temperature changes to bee losts. *Uludag Bee J* **4**, 171–174 (2004).
11. Seeley, T. Regulation of temperature in the nests of social insects. *Insect Thermoregulation*. pp. 159–234 (1981).
12. Seeley, T. Honeybee ecology. *Honeybee Ecology*. (2014).
13. Abou-Shaara, H., Owayss, A., Ibrahim, Y. & Basuny, N. A review of impacts of temperature and relative humidity on various activities of honey bees. *Insectes Sociaux*. **64** pp. 455–463 (2017).
14. Human, H., Nicolson, S. & Dietemann, V. Do honeybees, Apis mellifera scutellata, regulate humidity in their nest? *Naturwissenschaften*. **93** pp. 397–401 (2006).
15. Meikle, W., Rector, B., Mercadier, G. & Holst, N. Within-day variation in continuous hive weight data as a measure of honey bee colony activity. *Apidologie* **39**, 694–707 (2008).
16. Zacepins, A., Kviesis, A., Komasilovs, V. & Muhammad, F. Monitoring system for remote bee colony state detection. *Baltic Journal Of Modern Computing* **8**, 461–470 (2020).
17. Hunt, J. & Richard, F. Intracolony vibroacoustic communication in social insects. *Insectes Sociaux* **60**, 403–417 (2013).
18. Rustam, F., Sharif, M., Aljedaani, W., Lee, E. & Ashraf, I. Bee detection in bee hives using selective features from acoustic data. *Multimedia Tools And Applications* **83**, 23269–23296 (2024).
19. Ruvinga, S., Hunter, G., Duran, O. & Nebel, J. Use of LSTM Networks to Identify "Queenlessness" in Honeybee Hives from Audio Signals. *2021 17th International Conference On Intelligent Environments (IE)*. pp. 1–4 (2021).
20. Nolasco, I. *et al*. Audio-based identification of beehive states. *ICASSP 2019-2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 8256–8260 (2019).
21. Terenzi, A., Ortolani, N., Nolasco, I., Benetos, E. & Cecchi, S. Comparison of Feature Extraction Methods for Sound-based Classification of Honey Bee Activity. *IEEE/ACM Transactions On Audio, Speech, And Language Processing*. (2021).
22. Dubois, S. *et al*. Bee Detection For Fruit Cultivation. *2021 IEEE International Symposium On Circuits And Systems (ISCAS)*. pp. 1–5 (2021).
23. Nolasco, I. & Benetos, E. To bee or not to bee: Investigating machine learning approaches for beehive sound recognition. *ArXiv Preprint ArXiv:1811.06016*. (2018).
24. Zlatkova, A., Kokolanski, Z. & Tashkovski, D. Honeybees swarming detection approach by sound signal processing. *2020 XXIX International Scientific Conference Electronics (ET)*. pp. 1–3 (2020).
25. Žgank, A. Acoustic monitoring and classification of bee swarm activity using MFCC feature extraction and HMM acoustic modeling. *2018 ELEKTRO*. pp. 1–4 (2018).
26. Zlatkova, A., Gerazov, B., Tashkovski, D. & Kokolanski, Z. Analysis of parameters in algorithms for signal processing for swarming of honeybees. *2020 28th Telecommunications Forum (TELFOR)*. pp. 1–4 (2020).
27. Zhang, T., Zmyslony, S., Nozdrenkov, S., Smith, M. & Hopkins, B. Semi-Supervised Audio Representation Learning for Modeling Beehive Strengths. *ArXiv Preprint ArXiv:2105.10536*. (2021).
28. Qandour, A., Ahmad, I., Habibi, D. & Leppard, M. Remote Beehive Monitoring Using Acoustic Signals. *Acoustics Australia* **42**, 205 (2014).
29. Zhao, Y. *et al*. Based investigate of beehive sound to detect air pollutants by machine learning. *Ecological Informatics*. **61** pp. 101246 (2021).
30. Hunter, G., Howard, D., Gauvreau, S., Duran, O. & Busquets, R. Processing of multi-modal environmental signals recorded from a "smart" beehive. *Proceedings Of The Institute Of Acoustics* **41**, 339–348 (2019).
31. Zhu, Y. *et al*. Early prediction of honeybee hive winter survivability using multi-modal sensor data. *2023 IEEE International Workshop On Metrology For Agriculture And Forestry (MetroAgriFor)*. pp. – (2023).
32. Kulyukin, V. Audio, image, video, and weather datasets for continuous electronic beehive monitoring. *Applied Sciences* **11**, 4632 (2021).
33. Cecchi, S. *et al*. A preliminary study of sounds emitted by honey bees in a beehive. *Audio Engineering Society Convention 144*. (2018).
34. Zhu, Y. *et al*. MSPB: a longitudinal multi-sensor dataset with phenotypic trait measurements from honey bees. *Scientific Data* **11**, 860 (2024).
35. Chabert, S. *et al*. Rapid measurement of the adult worker population size in honey bees. *Ecological Indicators*. **122**, pp. 107313 (2021).
36. Abdollahi, M., Henry, E., Giovenazzo, P. & Falk, T. The Importance of Context Awareness in Acoustics-Based Automated Beehive Monitoring. *Applied Sciences* **13**, 195 (2022).
37. Di, N., Sharif, M., Hu, Z., Xue, R. & Yu, B. Applicability of VGGish embedding in bee colony monitoring: comparison with MFCC in colony sound classification. *PeerJ*. **11** pp. e14696 (2023).
38. Dietemann, V. *et al*. Standard methods for varroa research. *Journal Of Apicultural Research* **52**, 1–54 (2013).
39. Kohl, P. *et al*. Parasites, depredators, and limited resources as potential drivers of winter mortality of feral honeybee colonies in German forests. *Oecologia* **202**, 465–480 (2023).
40. Kempers, M. *et al*. STATEMENT ON HONEY BEE WINTERING LOSSES IN CANADA FOR 2023.
41. Abdollahi, M. *et al*. UrBAN: Urban Beehive Acoustics and PheNotyping Dataset. *Available*: https://doi.org/10.20383/103.0972 (2024).
42. Abdollahi, M., Coallier, N., Giovenazzo, P. & Falk, T. Performance Comparison of Voice Activity Detectors for Acoustic Beehive Monitoring. *2023 IEEE Canadian Conference On Electrical And Computer Engineering (CCECE)*. pp. 320–323 (2023).
43. Guimarães, H. *et al*. Adapting Self-Supervised Features for Background Speech Detection in Beehive Audio Recordings. *2023 IEEE International Workshop On Metrology For Agriculture And Forestry (MetroAgriFor)*. pp. 663–667 (2023).
44. Lin, L., Holmes, W. & Ambikairajah, E. Adaptive noise estimation algorithm for speech enhancement. *ELECTRONICS LETTERS-IEE* **39**, 754–754 (2003).
45. Davis, S. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions On Acoustics, Speech, And Signal Processing* **28**, 357–366 (1980).
46. Breiman, L. Random forests. *Machine Learning*. **45**, pp. 5–32 (2001).
47. Genuer, R., Poggi, J. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters* **31**, 2225–2236 (2010).
48. Maćkiewicz, A. & Ratajczak, W. Principal components analysis (PCA). *Computers & Geosciences* **19**, 303–342 (1993).

49. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions On Pattern Analysis And Machine Intelligence* **27**, 1226–1238 (2005).
50. Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. *Advances In Neural Information Processing Systems*. **30** (2017).
51. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N. & Kashino, K. BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation. *2021 International Joint Conference On Neural Networks (IJCNN)*. https://doi.org/10.1109/IJCNN52387.2021.9534474 (2021,7).
52. Yang, A. "Smart Bee Colony Monitor: Clips of Beehive Sounds," *Kaggle*. Available: https://www.kaggle.com/dsv/4451415. https://doi.org/10.34740/KAGGLE/DSV/4451415 (2022).

Acknowledgements
The authors acknowledge funding from NSERC via their Alliance program (ALLRP 548872-19), as well as Nectar Technologies Inc and the Centre de recherche en sciences animales de Deschambault for the support with data collection. The authors would like to thank Evan Henry for his valuable contributions to the dataset collection in the field.

Author contributions
M.A. contributed to the initial drafting of the manuscript and conducted the experiments in the technical validation. Y.Z. and H.G. contributed to the feedback to experimental results. N.C. contributed to sensor data acquisition and verified data records. S.M. contributed in editing and review of the manuscript. P.G. and T.F. contributed in conceptualizing the study and study design, and supervision of the study. All authors contributed to the critical revision of the manuscript. All authors had full access to all the data in the study and took responsibility for the decision to submit this draft for publication.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to T.H.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.