Improved Monitoring of Honey bee Colony Strength via Audio IoT Sensors, Modulation Tensorgrams and Recurrent Neural Networks

Mahsa Abdollahi, Yi Zhu, Heitor R. Guimarães, Nico Coallier, Ségolène Maucourt, Pierre Giovenazzo, and Tiago H. Falk

***Abstract*—Honey bees (*Apis mellifera*) play a crucial role in agriculture and ecosystem stability as key pollinators of crops and wild plants. As such, monitoring hive strength remotely with Internet of Things (IoT) sensors has become a crucial task. Previously, handcrafted features extracted from the modulation spectrum of audio IoT devices were shown to improve acoustic monitoring of colony strength. In this paper, we hypothesize that important discriminative information is present in the temporal dynamics of the modulation spectrum, but this information is discarded with prior methods. As such, we explore the use of a new modulation tensorgram where the time dimension is kept. This new representation is used as input to a convolutional neural network (CNN) and a convolutional recurrent deep neural networks (CRDNN). Using the public UrBAN dataset, which contains more than 3,000 hours of beehive audio recordings, we show that the proposed method improves both accuracy and cross-hive generalizability over prior benchmark methods, and the results further suggest improved robustness to noisy in-the-wild recording conditions. We use saliency maps and gradient-weighted class activation maps for explainability and show the importance of the modulation spectral temporal dynamics for the task at hand. Overall, our results suggest that accurate, generalizable, and robust acoustic monitoring of honey bee colony strength is possible.**

***Index Terms*—Beehive acoustics, honey bees, Modulation spectrogram.**

I. Introduction

Honey bees (Apis mellifera) play a vital role as pollinators in both agricultural and natural ecosystems, supporting the reproduction of crops and wild plants alike [1]. Despite their ecological and economic significance, global honey bee populations are declining at an alarming rate due to a combination of stressors, including pesticide use, parasitic mites, disease, and climate change [2], [3]. These losses pose a serious threat to biodiversity and food security [4]. Traditional hive assessments typically rely on manual inspections conducted at multi-week intervals, an approach that is not only time-intensive and laborious but can also be disruptive to the colony. Moreover, for commercial beekeepers managing hundreds or thousands of colonies, grading colony strength at scale using manual methods is practically infeasible. This creates significant management challenges because critical decisions, including mite treatment, supplemental feeding, and the selection of colonies for pollination contracts, depend heavily on accurate and timely assessments of colony strength. These constraints highlight the need for non-invasive and continuous monitoring systems capable of providing real-time and objective insights into hive health.

Advances in precision apiculture, particularly through the integration of Internet-of-Things (IoT) technologies, have enabled the development of automated systems for hive monitoring [5], [6]. These sensor-based platforms capture internal hive parameters such as temperature, humidity, and acoustic activity, transmitting data remotely to facilitate timely detection of critical events like pest infestation, queen loss, or behavioral abnormalities [7], [8]. Among these modalities, acoustic monitoring has emerged as especially valuable. Honey bees generate a rich variety of sounds via wingbeats, thoracic vibrations, and other movements, offering a non-invasive window into the colony health state [9]. In-hive microphones can record such buzzing sounds, which have been used to detect the presence of the queen bee [10], [11] and parasite infestations [7], [12], [13], detect swarming [14], predict winter survivability [15], and measure variations in colony strength [16], [17].

Beyond event detection, continuous acoustic monitoring also enables the observation of behavioral shifts within the colony. Because honey bees communicate primarily through pheromones and vibrational signals, stressors such as hornet attacks or Varroa mite infestations can significantly alter the acoustic profile of the hive [13], [18]. For example, defensive or alarm-related behaviors, including piping signals, may shift the distribution of frequencies toward higher ranges under predator or parasite pressure. Long-term audio monitoring further facilitates the characterization of circadian rhythms and daily activity patterns, providing deeper insights into colony dynamics and overall physiological state [19].

The performance of such systems depends heavily on the quality of the recorded audio data. Data quality has a direct impact on the features to be extracted and used for hive monitoring. Commonly, time-frequency based methods have been proposed and the developed monitoring tools have greatly relied on spectrogram [20], [21] and mel-frequency cepstral coefficient (MFCC) [11], [22] based features. MFCCs, for example, have been widely used in bee species recognition, swarming, and queen presence detection [23]. These time-frequency based audio methods, however, are very sensitive to overlapping noise sources, such as beekeeper speech, rain,

The authors acknowledge funding from NSERC via their Alliance program (ALLRP 548872-19), as well as Nectar Technologies Inc for the support with data collection.

M.A., Y. Z., H. G., and T. F. are with the INRS-EMT, Université du Québec, Montréal, Canada. (e-mail: mahsa.abdollahi@inrs.ca, Yi.Zhu@inrs.ca, Heitor.Guimaraes@inrs.ca, Tiago.Falk@inrs.ca).

S. M., P. G. are with the biology department, Laval university, Quebec, Canada. (e-mail: segolene.maucourt.1@ulaval.ca, pierre.giovenazzo@bio.ulaval.ca).

N. C. is with the Nectar Technologies Inc., Montréal, Canada. (e-mail: nico@nectar.buzz).

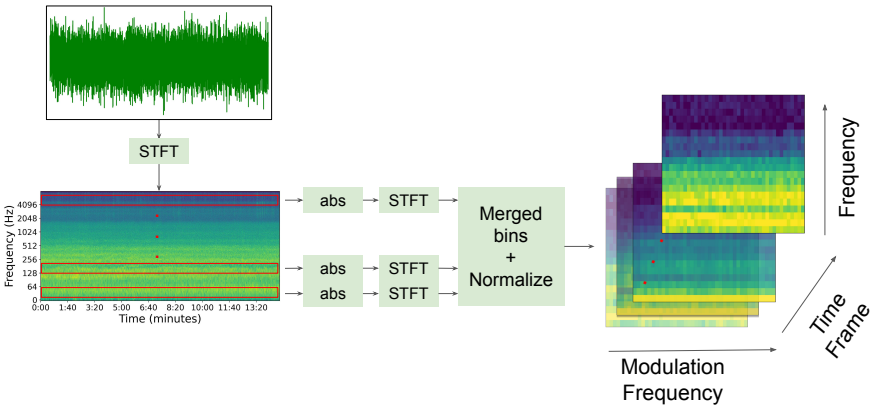


Fig. 1: Block diagram of signal processing steps involved in calculation of the modulation tensorgram.

wind, and other nearby factors, such as trains, tractors [17].

Moreover, while traditional hive monitoring tools relied on conventional machine learning classifiers [23], recently, there has been a rise on the use of deep neural networks, specifically convolutional neural networks (CNNs). CNNs were originally developed for image-like inputs, thus are well suited for spectrograms or mel-scaled spectrograms. The work in [24] employed a CNN with four convolutional layers to detect queenless hives using the spectrogram as input. In another study, CNNs were successfully applied to detect honey bee piping signals, a key indicator of colony reproductive dynamics [25]. Similarly, the work in [26] showed that CNN-based classifiers outperformed traditional machine learning models across several beehive monitoring tasks.

In parallel to these developments, the so-called modulation spectral features emerged to better capture the rhythmic and temporal patterns in beehive acoustics. These features highlight modulations in signal energy across time and frequency, making them well-suited for identifying the repetitive vibratory patterns typical of bee colonies and separating these patterns from unwanted noise sources and artifacts. In prior work, we proposed to handcraft some features from the modulation spectrum and we showed that such features were complementary to spectrogram-based ones using traditional machine learning classifiers, such as random forests [27]. Here, we build on this foundational work and explore the use of the raw modulation spectrogram as input to a CNNs and a convolutional recurrent deep neural networks (CRDNNs). This approach allows us to exploit the spatial and temporal richness of the modulation representations, as well as their noise-robustness properties, while leveraging the modeling power of deep neural networks. Our experiments show the proposed method generalizing to unseen conditions more efficiently than several benchmark methods, thus helping address a critical issue in remote beehive monitoring - generalizability across unseen environments [24], [28].

The novelty of the present study lies in three main aspects. First, we move beyond handcrafted modulation descriptors by directly using raw modulation spectrograms and time-preserving modulation tensorgrams as structured inputs to deep learning models, thereby preserving richer temporal–spectral information that is often lost in aggregated feature representations. Second, we provide a systematic evaluation of multiple deep learning architectures, including 2D CNNs, 3D CNNs, and CRDNNs, for the task of colony strength regression, allowing a comprehensive assessment of how different model capacities and inductive biases interact with modulation-based representations. Third, we incorporate explainability analyses using saliency maps and gradient-weighted class activation maps (Grad-CAM) visualizations to interpret the learned representations and to identify which time–frequency–modulation regions contribute most to the predictions, improving the transparency and interpretability of the proposed approach.

The remainder of this paper is organized as follows. Section II describes the proposed method. Section III describes the experimental setup, while Section IV discusses the obtained results. Lastly, Section V presents the conclusions.

II. Proposed Method

A. Modulation Spectrum

In many acoustic classification tasks, traditional spectrograms struggle in noisy environments. This limitation arises because both signal and noise are distributed across the same time-frequency space, making separation difficult. To address this limitation, the modulation spectrogram and its time-resolved counterpart, the modulation tensorgram, have been proposed [29]. Figure 1 depicts the signal processing steps involved in the computation of the modulation tensorgram.

First, a short-time Fourier transform (STFT) with a 100 ms window length and a 12.5 ms hop length is applied, resulting in the conventional spectrogram. Next, for each acoustic frequency bin independently, the spectral magnitude time series is extracted and a second STFT is computed across the time dimension using a 60 s window with no overlap, resulting in an intermediate tensor of $(freq \times modulation\ frequency \times time) = (801 \times 801 \times 14)$. These window and hop lengths were

optimized empirically based on pilot experimentation. For the short-term STFT, window lengths of 50 ms, 100 ms, and 200 ms were evaluated. Smaller windows (50 ms) provided higher temporal resolution but poorer frequency resolution, while larger windows (200 ms) improved frequency resolution at the expense of temporal precision and smoothed short acoustic fluctuations. A 100 ms window provided the best trade-off for capturing hive acoustic activity. Different hop sizes were also tested, since smaller hop lengths improve temporal continuity and allow higher observable modulation frequencies, whereas larger hop sizes reduce modulation resolution. A hop size of 12.5 ms yielded the most stable modulation representations.

For the long-term modulation analysis, aggregation windows ranging from 30 s to 5 min were explored. Shorter windows were more sensitive to transient environmental events, while larger windows excessively smoothed colony acoustic dynamics. A 60 s aggregation window provided the best balance between robustness and temporal specificity. Finally, no overlap was applied at the long-term aggregation stage to reduce computational redundancy and simplify downstream processing.

The resulting representation is a 3-dimensional tensor where the y-axis corresponds to acoustic frequency, the x-axis to modulation frequency (i.e., the rate of temporal variation within each acoustic frequency bin), and the z-axis to time, where each $freq \times modulation\ \ frequency$ slice corresponds to one second-stage STFT frame. This $freq \times modulation\ \ frequency \times time$ representation is referred to as a modulation tensorgram. Averaging across the time dimension yields the 2-dimensional modulation spectrogram representation.

Prior to bin merging, the acoustic-frequency axis was capped at 1000 Hz. This choice is motivated by honey bee bioacoustics: the wingbeat, thoracic, and buzzing sounds that characterize in-hive activity concentrate at low frequencies, with their fundamental and lower harmonic components falling well below 1000 Hz. The discriminative analysis presented later in Section II-B is consistent with this choice, with the most discriminative regions appearing in the lower portion of this range. To further reduce dimensionality, the retained acoustic-frequency bins were then grouped down to 20, and adjacent modulation-frequency bins were grouped to 40, resulting in a final tensor size of $20 \times 40 \times 14$ for the modulation tensorgram and 20×40 for the modulation spectrogram.

To illustrate some representative modulation spectrograms, Figure 2 (a)-(c) shows an average representation for beehives of varying strengths, namely with frames of bees (fob) ranging from $1 \leq$ *fobs* ≤ 10, $11 \leq$ *fobs* ≤ 20, and $21 \leq$ *fobs* ≤ 30, respectively. For visualization purposes, these examples were selected from recordings collected between 8 pm and 11 pm, a period corresponding to peak hive occupancy when the majority of forager bees have returned to the colony. This time window provides acoustic patterns that are more representative of the collective colony state and less influenced by transient variability associated with daytime foraging activity.

B. Discriminative power of the modulation spectrum

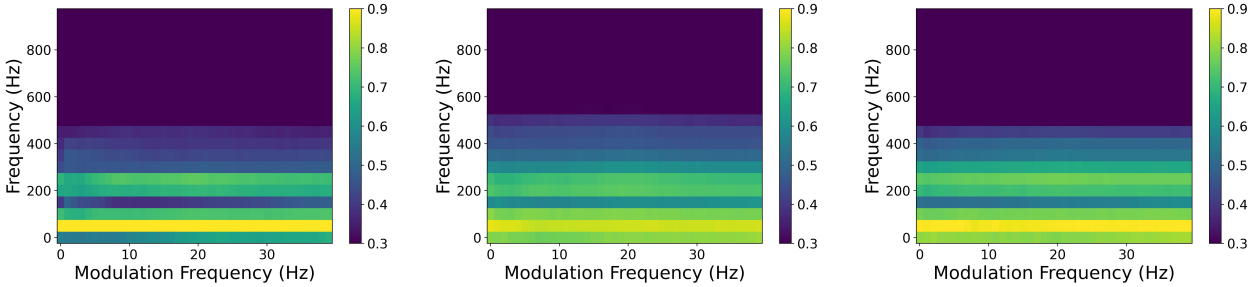
To measure the discriminative power of the modulation spectrogram, we compute the Fisher ratio (F-ratio) between different hive strength groups. The F-ratio quantifies how well a given feature distinguishes between two classes by comparing two types of variance: the between-group variance, which captures the separation between group means, and the within-group variance, which reflects variability within each group. A higher F-ratio value indicates greater discriminability.

To compute the F-ratio, the number of frames of bees (*fobs*) was divided into three population categories: Category 1 includes $1 \leq$ *fobs* ≤ 10, Category 2 includes $11 \leq$ *fobs* ≤ 20, and Category 3 includes $21 \leq$ *fobs* ≤ 30. These three categories correspond to the conditions shown in Figures 2 (a)-(c). Figure 3 presents the resulting F-ratio maps for the following pairwise comparisons: (a) Category 1 vs. Category 2, (b) Category 2 vs. Category 3, and (c) Category 1 vs. Category 3. These maps highlight the modulation–spectral regions where the differences in energy are most pronounced across varying levels of bee activity. In particular, the lower-frequency regions around 150–200 Hz, as well as the band near 500–600 Hz, show the strongest discriminative power, indicating that changes in these modulation components should be closely associated with increases in colony activity and could be leveraged for automated beehive strength monitoring tasks.

C. Deep neural network model architectures

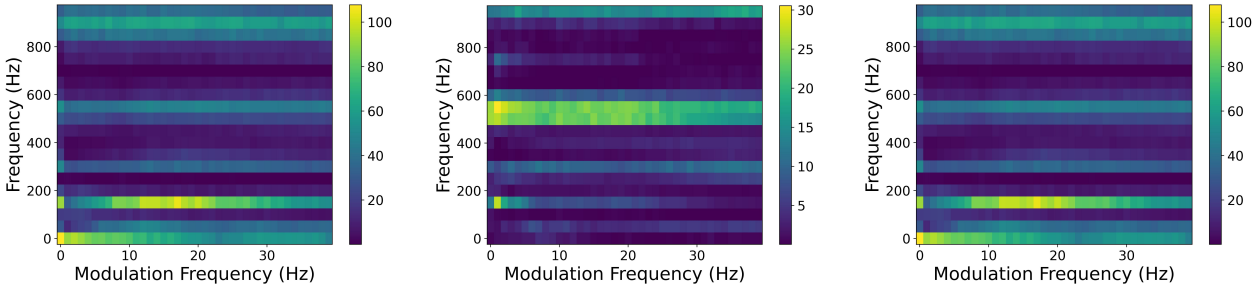
To model the acoustic patterns of beehive audio, we implemented several deep learning architectures capable of learning from 2D and 3D input representations. These models are optimized to handle spectrograms, modulation spectrograms, and modulation tensorgrams. The architectures tested include a baseline CNN, a CNN with spatial attention, a convolutional-recurrent deep neural network (CRDNN), and a 3D variant for modulation tensorgrams. Models were trained using the Keras framework [30] for 300 epochs using the Adam optimizer. All models were trained using the RMSE as the loss function. An initial learning rate of 0.0001 was employed, together with an adaptive learning rate strategy to improve convergence during training. Experiments were conducted with batch sizes of 128. Early stopping was used to monitor the validation loss and terminate training if no meaningful improvement (minimum change of 1×10^{-4}) was observed over 15 consecutive epochs. The best-performing model weights were restored at the end of training. In addition, a ReduceLROnPlateau strategy was applied to adaptively decrease the learning rate by a factor of 0.5 when the validation loss plateaued for 5 epochs, with a lower bound of 1×10^{-9}, ensuring continued but controlled learning. More details on these architectures and benchmark systems is given next.

1) Baseline CNN: The baseline convolutional model consists of three convolutional layers using 3×3 kernels and ReLU activation. Each block is followed by batch normalization and pooling, with L1 regularization applied to the deeper layers to reduce overfitting. The final convolutional features are flattened and passed through a dense layer with 64 units,



(a) (b) (c)

Fig. 2: Modulation spectrograms averaged on samples with a) $1 \leq fobs \leq 10$, b) $11 \leq fobs \leq 20$, and c) $21 \leq fobs \leq 30$.



(a) (b) (c)

Fig. 3: The number of frames of bees (*fobs*) is grouped into three categories: Category 1 corresponds to $1 \leq fobs \leq 10$, Category 2 to $11 \leq fobs \leq 20$, and Category 3 to $21 \leq fobs \leq 30$. F-ratio maps comparing the following pairs of categories: a) Category 1 vs. Category 2, b) Category 2 vs. Category 3, and c) Category 1 vs. Category 3.

TABLE I: Details of baseline CNN model

Layer	**Details**
Conv2D	32 filters, 3×3, ReLU, same padding
BatchNorm	–
MaxPool2D	2×2 pool
Conv2D	64 filters, 3×3, ReLU, L1 reg, same padding
BatchNorm	–
MaxPool2D	2×2 pool
Conv2D	128 filters, 3×3, ReLU, L1 reg, same padding
BatchNorm	–
Dropout	Rate 0.25
Flatten	–
Dense	64 units, ReLU
BatchNorm	–
Dropout	Rate 0.5
Dense	1 unit, linear activation

TABLE II: Details of CNN model with spatial attention

Layer	**Details**
Conv2D	32 filters, 3×3, ReLU, same padding
BatchNorm	–
Spatial Attn.	7x7 Conv2D, sigmoid
MaxPool2D	2×2 pool
Conv2D	64 filters, 3×3, ReLU, L2 reg
BatchNorm	–
Spatial Attn.	7x7 Conv2D, sigmoid
MaxPool2D	2×2 pool
Conv2D	128 filters, 3×3, ReLU, L2 reg
BatchNorm	–
Spatial Attn.	7x7 Conv2D, sigmoid
Dropout	Rate 0.25
Flatten	–
Dense	64 units, ReLU
BatchNorm	–
Dropout	Rate 0.5
Dense	1 unit, linear activation

followed by dropout, and finally mapped to a scalar regression output. Details about this architecture are given in Table I.

2) CNN with Spatial Attention: To enhance model focus on salient acoustic regions, spatial attention modules were integrated into the CNN pipeline (see Table II for more details). Attention maps are computed using a 7×7 convolution after each convolutional layer, emphasizing relevant feature regions. This attention-guided feature enhancement improves robustness to irrelevant background noise, while L2 regularization and dropout are employed to reduce overfitting.

3) CRDNN (Convolutional-Recurrent-DNN): To capture both local and long-range temporal dependencies in the acoustic data, we designed a CRDNN architecture (see Table III). The model begins with three convolutional layers, each followed by batch normalization and pooling. After flattening, the output is reshaped into a temporal sequence suitable for a Gated Recurrent Unit (GRU) processing. A two-layer GRU extracts sequential patterns, followed by fully connected layers for final prediction. This hybrid structure combines

TABLE III: Details of CRDNN model

Layer	Details
Conv2D	32 filters, 3×3, ReLU, same padding
BatchNorm	–
MaxPool2D	2×2 pool
Conv2D	64 filters, 3×3, ReLU, L1 reg, same padding
BatchNorm	–
MaxPool2D	2×2 pool
Conv2D	128 filters, 3×3, ReLU, L1 reg, same padding
BatchNorm	–
Dropout	Rate 0.25
Flatten	–
Reshape	To $(\text{seq\_len}, 128)$
GRU	32 units, return seq = True
Dropout	Rate 0.5
GRU	64 units, return seq = False
Dense	256 units, ReLU
BatchNorm	–
Dropout	Rate 0.5
Dense	128 units, ReLU
BatchNorm	–
Dropout	Rate 0.5
Dense	1 unit, linear activation

the strengths of CNNs for spatial and GRUs for temporal modeling.

4) 3D variants: Lastly, to leverage the full dimensionality of the modulation tensorgram representations, we designed 3D variants of our convolutional and attention-based models. Unlike the 2D models, which operate on standard time-frequency spectrograms, the 3D models process volumetric inputs that capture time, frequency, and modulation frequency simultaneously. This three-dimensional structure enables the models to extract joint spatio-spectro-temporal features, making them well-suited for tasks involving modulation-based acoustic representations. The 3D CNN with attention integrates spatial attention mechanisms across the 3D input, enhancing relevant regions within the tensor while suppressing less informative ones. Each convolutional block includes batch normalization, 3D max-pooling, and dropout for regularization, followed by dense layers for regression output. By capturing patterns across all axes of the tensorgram, these models are able to exploit modulation structures more effectively, improving the robustness and generalizability of acoustic monitoring tasks.

III. Experimental Setup

A. Dataset Description

To test the efficacy of our proposed method, we rely on the UrBAN dataset [31], which includes over 3,000 hours of raw audio data collected from an urban apiary. The UrBAN dataset is comprised of audio data collected from nine beehives monitored over a two-year period at a rooftop apiary in Montréal, Québec, Canada. Each hive included one brood chamber and one-to-two honey supers, housed within 10-frame standard Langstroth boxes.

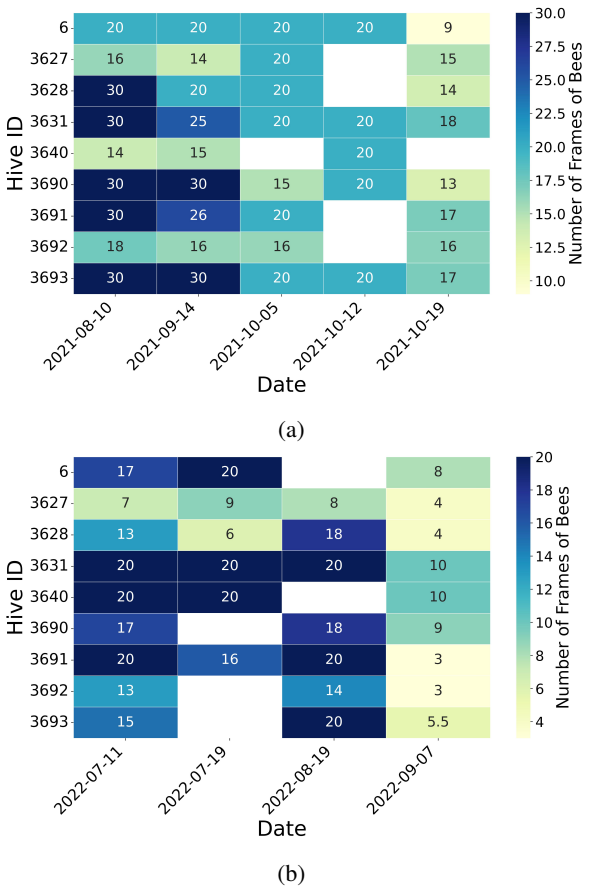


Fig. 4: Heatmaps showing the number of frames of bees estimated during visual inspection for each hive throughout (a) 2021 and (b) 2022. Values between 1-10 suggest only the brood chamber was present, 11-20 brood chamber plus one super, and 21-30 brood chamber plus two supers.

The hives were manually inspected to assess hive strength, verify the presence of a laying queen, and record any additional colony activity observations. Hive strength is defined as the number of frames where at least 70% of the frame is covered by bees [32]. A microphone was installed atop the central frame of the bottom brood box. The audio data consists of 15-minute audio segments recorded every 30 minutes at a sampling rate of 16 kHz. Figure 4 shows the number of frames of bees (fob) for each beehive in 2021 (top plot) and 2022 (bottom). Values between 1-10 suggest only the brood chamber was present, while 11-20 the brood chamber plus one super, and 21-30 the brood chamber plus two supers.

Because manual inspections occur at discrete dates while audio is recorded continuously, a colony strength label was assigned to each 15-minute audio segment by linear temporal interpolation between consecutive inspection measurements rather than by nearest-neighbour assignment. This procedure yields temporally continuous targets that better reflect the gradual population dynamics between inspections.

B. Benchmark features and noise removal

To establish a baseline for evaluating the performance of the proposed features, we incorporated two widely recognized acoustic benchmarks into our analysis.

MFCCs: These features are designed to simulate the non-linear human auditory system by applying a mel-scale frequency mapping, making them highly effective for bioacoustic tasks. MFCCs are a standard benchmark in precision beekeeping and have proven valuable in characterizing bee sounds [23]. For this study, we extracted 13 coefficients (12 standard coefficients plus the zeroth coefficient, which represents the signal's energy) using 26 mel-filter banks. The resulting MFCC representation has dimensions of 13 × 40, where the y-axis corresponds to the 13 cepstral coefficients and the x-axis to 40 time frames, matching the time-axis resolution of the modulation representations.

Spectrograms: As a fundamental representation of audio signals, spectrograms visualize the evolution of spectral content over time (time-frequency analysis). They are extensively used in beehive acoustics, notably for critical tasks such as queen bee detection [10] and swarming prediction [20], making them an essential comparison point for any new monitoring method. Consistent with the modulation representations, the spectrogram frequency axis was capped at 1000 Hz, and the resulting frequency bins were grouped down to 20 bins. This yields a final spectrogram of dimension 20×40 ($frequency \times time$), matching the shape of the modulation spectrogram.

Notwithstanding, as mentioned previously, these features have been shown to be sensitive to environmental noise. As the dataset used herein was collected "in the wild," it is crucial that some noise suppression be performed to ensure reliable prediction from the benchmark systems. Here, we employ a classical spectral amplitude subtraction method for noise reduction, as described in [31].

C. Testing Setup and Figures-of-Merit

Building on recommendations from previous studies, we explored two distinct experimental configurations, namely "random-split" and "hive-independent".

In the random-split approach, partitioning was performed at the level of individual 15-minute audio segments, where samples were randomly assigned to the training, validation, and testing folds while preserving the overall label distribution. A nested 5-fold cross-validation strategy was then employed to ensure robust evaluation of the models. This procedure ensured that every data point was included in the test set exactly once, reducing the risk of overfitting and providing a more reliable estimate of model performance. However, because temporally adjacent recordings may exhibit strong acoustic correlations, this setup may produce optimistic performance estimates due to similarities between training and testing samples collected from nearby time periods.

Conversely, in the hive-independent setup, the folds were constructed at the hive level such that all recordings originating from a given hive were assigned exclusively to either the training, validation, or testing subsets. This prevented any acoustic information from the same colony from appearing across splits. The 5 folds were therefore selected hive-independently for both validation and testing. This latter setup is more stringent and more closely resembles real-world scenarios where models must generalize to unseen hives and environmental conditions. As such, it provides a more realistic assessment of model robustness and generalizability to unseen environments.

As our task is a regression aimed at predicting the strength of a hive (given by its fob value), two conventional figures-of-merit are utilized: mean absolute error (MAE) and Pearson correlation ($\boldsymbol{r}$) between the real and predicted fob values. An ideal/perfect regressor will have MAE close to zero and a correlation close to unity.

IV. Results and Discussion

A. Colony Population Strength Prediction

Table IV presents the performance achieved by the different benchmark and proposed methods. As can be seen, in the random-split scenario, while the spectrogram and MFCC inputs yield MAE values around 3.5 and correlation $r \approx 0.76$, the modulation-based methods are capable of reducing the MAE to around 1.5 and elevate the correlation to $r \approx 0.95$. This underscores the hypothesis that bee acoustic activity is best characterized by temporal modulations (i.e., the rate and structure of changes in the dominant frequencies) rather than the static spectral envelope or cepstral coefficients. Moreover, the traditional spectrogram and MFCC benchmarks showed substantial gains in performance once audio enhancement was applied. The gains brought by spectral subtraction were noticeably smaller for the proposed method than for the spectrogram and MFCC baselines. This pattern is consistent with the previously reported noise-robustness of modulation spectral representations [27]. We note, however, that the UrBAN dataset does not include per-segment annotations of environmental disturbances (rain, wind, vehicles, etc.), so this observation suggests — rather than directly demonstrates improved robustness; a stratified evaluation against labelled noise events is left as future work (see Section IV-F).

Moreover, some attention-based variants do not improve over, and in a few cases trail, their non-attention counterparts after enhancement (for example Spec-CNN-att-2D vs. Spec-CNN-2D in the random-split setting). We attribute this to two effects: (i) once spectral subtraction has removed much of the broadband background, the residual input contains less irrelevant content for spatial attention to suppress; and (ii) the local musical-noise artifacts that classical spectral subtraction can introduce in low-energy regions are high-contrast and may be incidentally selected by the attention maps. Consistent with this interpretation, attention continues to provide a benefit on the more noise-resilient modulation-based features and on the more challenging hive-independent setting. Overall, for this setting, the CRDNN-3D model utilizing the modulation tensorgram as input resulted in the best performance.

For the more strict hive-independent setting, the performance is shown to drop for all models and features. Notwithstanding, overall, the modulation spectrogram and tensorgram inputs retain their superior performance, achieving correlations

greater than 0.7, while the spectrogram and MFCCs benchmarks remain around 0.50. This resilience suggests that the underlying frequency-modulation patterns are more robust and generalizable indicators of bee activity than the raw spectral content, which is likely dominated by hive-specific artifacts.

The CRDNN-3D achieved the lowest MAE in both scenarios (1.01 and 3.31). In the random-split setting, where the standard deviations are tight (e.g., 1.01 ± 0.08), this advantage is well separated from the competing methods. In the hive-independent setting the numerical advantage falls within the range of variability, so we motivate CRDNN-3D as our recommended model not on the basis of this margin alone, but on its consistently strong average ranks across the Friedman procedure (Table VI) and its favorable accuracy–efficiency–size tradeoff (Section IV-E).

B. Non-parametric Statistical Analyses

To statistically evaluate the differences in performance seen amongst the models, we employed the Friedman test [33] across all four experimental scenarios (random-split and hive-independent, with and without audio enhancement) and two evaluation metrics. Next, we used the Nemenyi post-hoc test [34] to conduct pairwise comparisons and identify which specific differences were statistically significant. The results of the Friedman test are summarized in Table V and showed statistically significant differences ($p < 0.05$) among the feature set and models performances across evaluation metrics. The critical Chi-square value for this test (with $k = 11$) was determined to be 18.30 at $\alpha = 0.05$. Next, Table VI reports the average ranks derived from the Friedman procedure. For error metrics such as MAE, a lower rank signifies superior performance, while for predictive metrics such as Correlation, a higher rank is better. Notably, the CRDNN-3D achieved particularly strong average ranks, especially when tested under the challenging hive-independent scenario.

Figures 5 (a)-(b) present the Critical Difference (CD) diagrams for the MAE metric under both the random-split and hive-independent evaluation setups, respectively. These diagrams summarize the post-hoc statistical comparisons: methods linked by a horizontal bar do not differ significantly, while those separated by more than the critical distance show statistically significant differences. As can be seen, in the random-split scenario, Mod-Tgram(tensorgram)-CRDNN-3D obtained the top average rank and outperformed models such as MFCCs-CNN-2D and Spec-CNN-att-2D, evidenced by the clear separation in the CD diagram. Conversely, in the hive-independent scenario, Mod-Tgram-CRDNN-3D grouped distinctly from MFCCs-att-CNN-2D, highlighting its superior ability to generalize across hives. Several of the modulation-based methods fall within the same statistical group, so the apparent ordering within that group should be interpreted as comparable performance rather than a strict ranking.

C. Explainability analysis

To interpret the spatial and temporal relevance of the input modulation spectrograms, we generated saliency maps for the trained 2D CNN and attention-based models. Saliency maps highlight the input regions that most strongly influence the model's predictions by computing the gradient of the output with respect to each feature. In our case, these maps reveal which frequency and modulation bands are most discriminative for different levels of bee activity. To reduce noise and capture consistent activation patterns, we averaged the saliency maps across multiple samples within each colony strength category (low, medium, and high). The resulting maps indicate that specific modulation frequency regions contribute more strongly as colony strength increases, suggesting that the model relies on structured acoustic patterns rather than isolated features. Figures 6 and 7 show the saliency maps for the three strength levels for the 2D CNN and 2D CNN with spatial attention models, respectively.

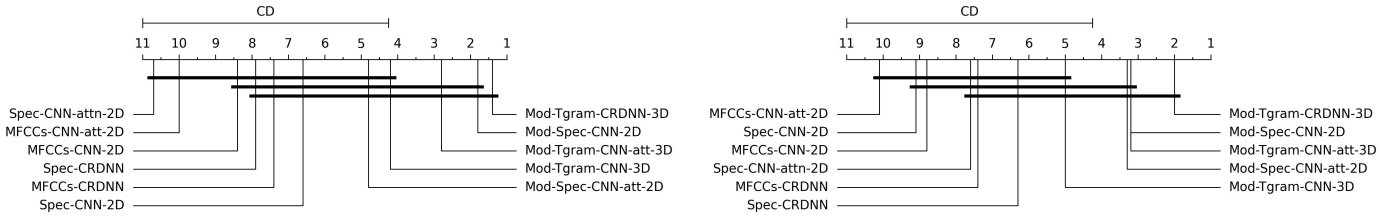
The saliency maps for both the standard 2D CNN and the 2D CNN with spatial attention confirm that the models primarily rely on specific modulation-frequency patterns to predict bee activity levels, suggesting a reliance on structured acoustic patterns. For the standard 2D CNN (Figure 6), increasing colony strength correlates with a progressive expansion and upward shift of the relevant frequency region, suggesting that higher harmonics or broader spectral features become relevant. In contrast, the 2D CNN with spatial attention (Figure 7) maintains a consistently narrower, more horizontal high-saliency band, primarily focused around 200 Hz, which expands across the modulation frequency axis as activity increases. This suggests the attention mechanism is effective at isolating the single most discriminative fundamental frequency band and focusing on its temporal variation (modulation), whereas the standard CNN integrates a broader range of adjacent frequencies. Notably, both models show a peak relevance in the 10–25 Hz modulation frequency range across the medium and high strength categories, indicating that these specific rates of temporal fluctuation are crucial discriminative features.

Next, for explainability of the 3D CNN model, we used the gradient-weighted class activation mapping (Grad-CAM) method [35]. Grad-CAM leverages intermediate feature activations to generate interpretable maps across spatial and temporal dimensions. By averaging these maps across samples, we identified the regions most relevant to the model's predictions, revealing how specific frequency–modulation patterns evolve with colony strength. This approach allowed us to visualize and compare differences in the 2D and 3D architectures.

Rather than focusing solely on a single aggregated explanation, we examined several representative frames including the first, middle, and last frames to capture the model's response to temporal variations in the modulation spectrogram. We found that high-activity samples exhibit more consistent and spatially distributed activation across frames, while low-activity samples display more localized and intermittent patterns. These observations suggest that the 3D models leverage both spectral structure and temporal evolution to estimate colony strength levels. Figure 8 provides a detailed visualization of these temporal dynamics. For each strength level (left column=low, middle=medium, right=high), the figure shows the three Grad-CAM heatmaps extracted from three different time frames (top row=first, middle=middle, bottom=last frame).

Feature	Model	Random-Split			
		No pre-processing		Spectral amplitude subtraction	
		MAE	r	MAE	r
Spectrogram	CNN-2D	3.70 ± 0.17	0.76 ± 0.03	1.71 ± 0.13	0.83 ± 0.02
	CNN-att-2D	3.68 ± 0.08	0.75 ± 0.02	2.17 ± 0.09	0.76 ± 0.01
	CRDNN-2D	3.49 ± 0.24	0.78 ± 0.02	1.77 ± 0.05	0.81 ± 0.01
MFCCs	CNN-2D	3.62 ± 0.13	0.76 ± 0.02	1.85 ± 0.25	0.81 ± 0.03
	CNN-att-2D	3.72 ± 0.09	0.75 ± 0.01	2.15 ± 0.17	0.77 ± 0.01
	CRDNN-2D	3.37 ± 0.05	0.78 ± 0.03	1.78 ± 0.10	0.82 ± 0.01
Modulation spectrogram	CNN-2D	1.54 ± 0.14	0.95 ± 0.01	1.14 ± 0.06	0.96 ± 0.01
	CNN-att-2D	1.52 ± 0.16	0.96 ± 0.03	1.39 ± 0.06	0.97 ± 0.03
Modulation tensorgram	CNN-3D	1.61 ± 0.10	0.94 ± 0.02	1.36 ± 0.09	0.96 ± 0.01
	CNN-att-3D	1.45 ± 0.05	0.96 ± 0.02	1.23 ± 0.06	0.96 ± 0.03
	CRDNN-3D	1.24 ± 0.03	0.95 ± 0.01	1.01 ± 0.08	0.97 ± 0.01
		Hive-Independent			
Spectrogram	CNN-2D	4.87 ± 1.47	0.47 ± 0.22	4.74 ± 1.46	0.52 ± 0.21
	CNN-att-2D	4.60 ± 1.61	0.46 ± 0.21	4.50 ± 1.59	0.51± 0.20
	CRDNN-2D	4.55 ± 1.62	0.56 ± 0.20	4.22 ± 1.16	0.61 ± 0.20
MFCCs	CNN-2D	4.95 ± 1.37	0.50 ± 0.22	4.65 ± 1.32	0.53 ± 0.22
	CNN-att-2D	4.99 ± 1.45	0.47 ± 0.24	4.97 ± 1.46	0.50 ± 0.22
	CRDNN-2D	4.42 ± 1.68	0.52 ± 0.25	4.38 ± 1.68	0.54 ± 0.20
Modulation spectrogram	CNN-2D	3.48 ± 0.90	0.69 ± 0.13	3.41 ± 0.87	0.74 ± 0.11
	CNN-att-2D	3.66 ± 1.09	0.72 ± 0.09	3.56 ± 1.05	0.76 ± 0.08
Modulation tensorgram	CNN-3D	3.84 ± 1.21	0.65 ± 0.22	3.69 ± 1.60	0.69 ± 0.21
	CNN-att-3D	3.76 ± 1.56	0.71 ± 0.15	3.67 ± 1.55	0.76 ± 0.16
	CRDNN-3D	3.42 ± 1.37	0.72 ± 0.20	3.31 ± 1.36	0.78 ± 0.17

TABLE IV: Performance comparison of MAE and correlation (r) between different feature sets and data partitioning setups, with and without spectral enhancement.



(a) (b)

Fig. 5: Critical Difference (CD) plot illustrating the results of the Nemenyi post-hoc test following the Friedman test on the MAE for a) random-split and b) hive-independent scenario. The models and features combination are ranked by their average performance across the five folds. Features grouped by a thick horizontal bar exhibit no statistically significant difference in performance ($\alpha = 0.05$). Features with lower ranks (towards 1) demonstrate superior performance.

As can be seen, for low strength colonies (1–10 fobs), the model's focus is generally localized to low frequencies (below 400 Hz) and very low modulation frequencies (below 10 Hz), becoming notably intermittent and contracting severely at the middle frame (t_{mid}). In contrast, high strength colony (21–30 fobs) samples show a consistent and broadly distributed activation across all frames, leveraging low-to-mid frequencies (up to 550 Hz) and spanning the entire range of modulation frequencies (up to 35 Hz), which suggests the model uses a stable, widespread feature set. The medium strength class (11–20 fobs) exhibits a dynamic pattern, shifting from a broad focus in t_1 to a localized one in t_{mid}, and then expanding again in $t_{\max}$. These observations confirm that the 3D architecture effectively leverages both spectral structure (Frequency) and its temporal evolution (Modulation Frequency over time) to differentiate colony strength levels.

D. Deep learning vs. Conventional classifiers

As a last exploration, we wish to compare the results obtained with the deep learning methods described herein with those of our earlier work that relied on a conventional random forest (RF) classifier [27]. Table VII summarizes

Metric	Random-Split			
	No pre-processing		Spectral amplitude subtraction	
	Friedman χ^2	p-value	Friedman χ^2	p-value
MAE	46.32	1.25e-06	47.30	8.31e-07
r	46.98	9.48e-07	47.44	7.82e-07
	Hive-Independent			
MAE	32.78	2e-04	35.90	8.75e-05
r	37.71	4.24e-05	40.32	1.48e-05

TABLE V: Friedman test results across evaluation metrics MAE and correlation(r) under random-split and hive-independent scenarios.

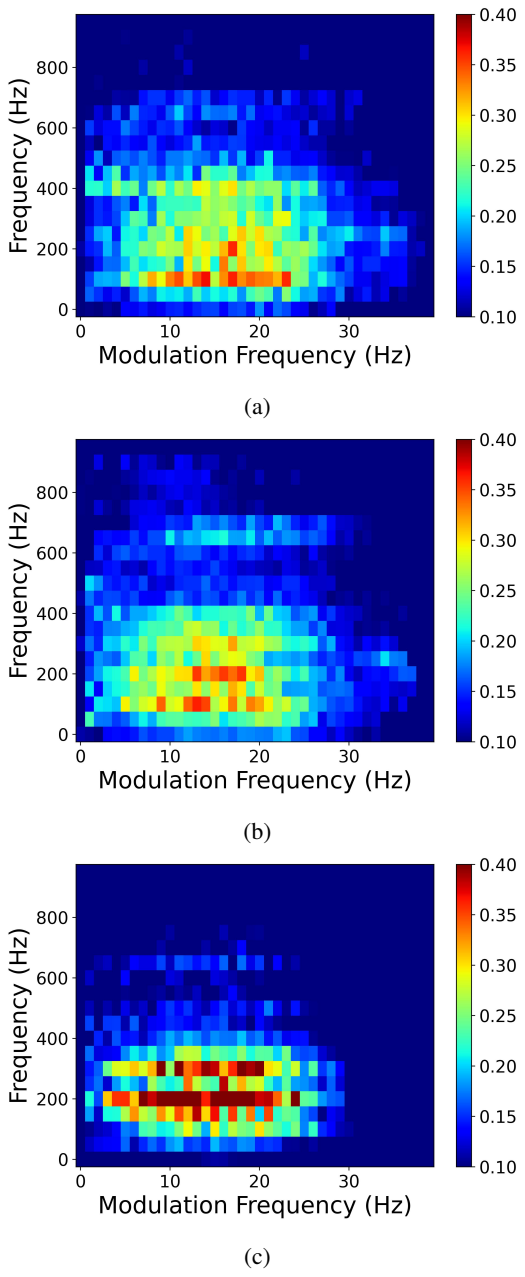
Feature	Model	Random-Split			
		No pre-processing		Spectral amplitude subtraction	
		MAE ↓	r↑	MAE ↓	r↑
Spectrogram	CNN-2D	6.6	4.0	6.6	5.5
	CNN-att-2D	10.7	2.8	10.7	1.3
	CRDNN-2D	7.9	3.6	7.9	3.7
MFCCs	CNN-2D	8.4	2.0	8.4	3.9
	CNN-att-2D	10.0	2.8	10.0	1.7
	CRDNN-2D	7.4	5.8	7.4	4.9
Modulation spectrogram	CNN-2D	1.8	7.9	1.8	9.9
	CNN-att-2D	4.8	9.4	4.8	8.16
Modulation tensorgram	CNN-3D	4.2	7.9	4.2	7.9
	CNN-att-3D	2.8	9.9	2.8	9.0
	CRDNN-3D	**1.4**	**9.9**	**1.4**	**10.1**
		Hive-Independent			
Spectrogram	CNN-2D	9.3	2.5	9.1	3.4
	CNN-att-2D	7.6	2.6	7.6	3.1
	CRDNN-2D	7.2	6.0	6.3	6.4
MFCCs	CNN-2D	9.2	4.1	8.8	3.0
	CNN-att-2D	9.1	3.0	10.1	2.4
	CRDNN-2D	6.6	3.8	7.4	3.2
Modulation spectrogram	CNN-2D	3.2	8.2	3.2	9.1
	CNN-att-2D	3.8	8.7	3.3	8.6
Modulation tensorgram	CNN-3D	5.1	8.1	5.0	8.2
	CNN-att-3D	2.9	9.7	3.2	9.4
	CRDNN-3D	**2.0**	**9.3**	**2.0**	**9.2**

TABLE VI: Comparison of average ranks across evaluation metrics under random-split and hive-independent scenarios, with and without audio enhancement. Bold values correspond to best methods.

the results obtained with the best RF model and ours. As can be seen, the proposed method improves accuracy across both figures-of-merit in the more stringent hive-independent scenario. Particularly, the proposed CRDNN-3D model is able to take advantage of the modulation tensorgram 3D input to better characterize the temporal changes of the bee buzzing sounds, something that the RF classifier with handcrafted features is not able to achieve.

E. Computational Complexity

The computational complexity results, summarized in Table VIII, were obtained on a MacBook Pro equipped with an Apple M1 chip and 16 GB of unified memory. Model sizes



(a)

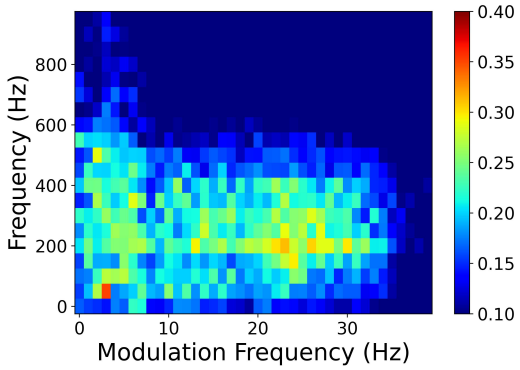
(b)

(c)

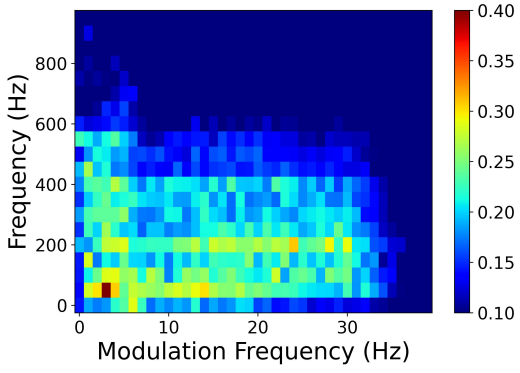
Fig. 6: Saliency maps of the trained 2D CNN model averaged over samples for each colony strength level: (a) low (1–10 fobs), (b) medium (11-20 fobs), and (c) high (21–30 fobs).

TABLE VII: Performance comparison of best-performing random forest (RF) classifier and best-performing deep learning (DL) method presented herein in the hive-independent scenario.

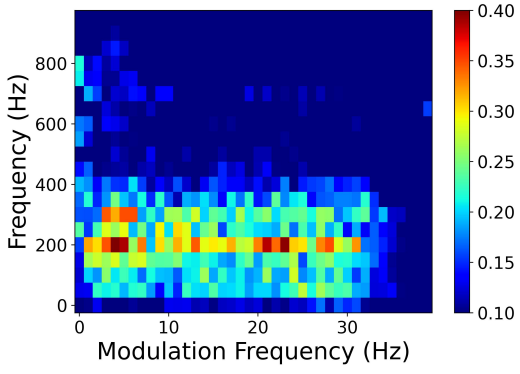
Metric	**RF Best (MSAB)**	**DL Best (CRDNN-3D)**
MAE	3.41 ± 1.03	3.31 ± 1.36
r	0.74 ± 0.15	0.78 ± 0.17



(a)



(b)



(c)

Fig. 7: Saliency maps of the trained 2D CNN model with spatial attention averaged over samples for each colony strength level: (a) low (1–10 fobs), (b) medium (11–20 fobs), and (c) high (21–30 fobs).

were extracted directly from the saved .h5 files, while inference times correspond to the average runtime over multiple forward passes per model.

Overall, 2D CNN-based models offer the most efficient inference, with CNN-2D achieving the lowest runtime (0.29 ms) and a compact footprint (6.1 MB). The addition of spatial attention slightly increases computational cost (0.41 ms), while remaining lightweight and suitable for real-time IoT deployment. 3D architectures are more computationally demanding due to volumetric feature extraction. CNN-3D and CNN-att-3D exhibit higher inference times (3.93 ms and 10.70 ms, respectively) and larger model sizes, reflecting the added cost of 3D convolutions and attention mechanisms.

TABLE VIII: Computational complexity of the evaluated deep learning models.

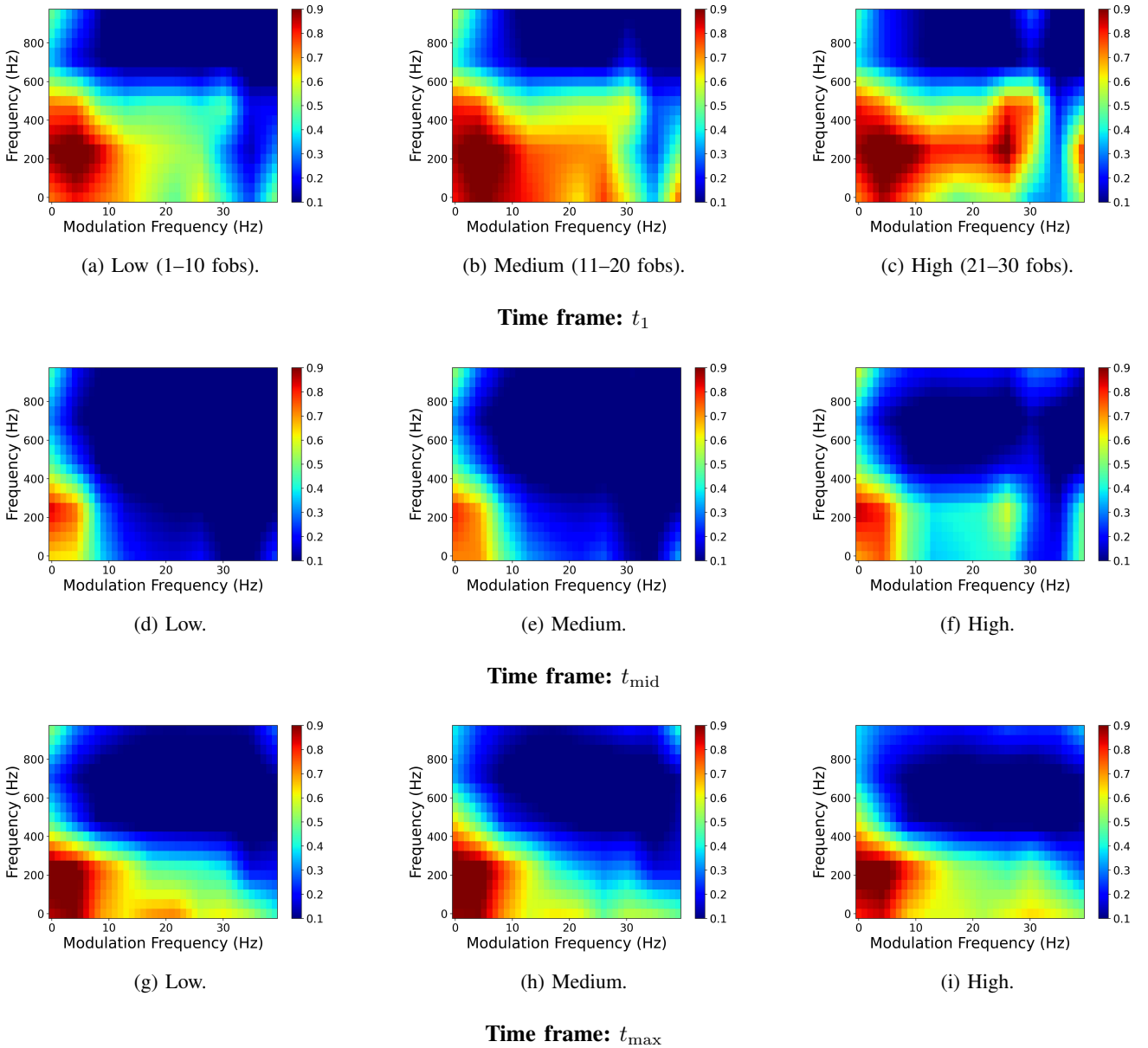
Model	# Parameters	Model Size (MB)	Inference Time (ms)
CNN-2D	503,553	6.1	0.29
CNN-att-2D	503,850	6.2	0.41
CRDNN-2D	179,137	2.3	0.62
CNN-3D	1,507,649	18.2	3.93
CNN-att-3D	1,919,310	23.2	10.70
CRDNN-3D	364,033	4.5	4.71

In contrast, CRDNN models provide a favorable balance between efficiency and representational power. Although CRDNN-3D has a higher inference time than 2D CNNs (4.71 ms), it remains significantly more efficient than CNN-att-3D while achieving the best overall predictive performance. This suggests that the recurrent component effectively captures temporal dependencies in the data, leading to improved accuracy without an excessive increase in model complexity. Notably, CRDNN-3D also maintains a relatively moderate model size (4.5 MB), making it a strong candidate for practical deployment. Models with size below 10 MB and inference under 5 ms on a consumer CPU (CRDNN-2D, CNN-2D, CNN-att-2D, CRDNN-3D) are plausible candidates for on-hive edge deployment; the heavier 3D CNNs (CNN-3D, CNN-att-3D) are more naturally deployed cloud-side with the IoT sensor uploading raw audio.

F. Limitations and Future Directions

One limitation of the present study is that colony strength may correlate with several external and hive-specific factors beyond the intrinsic acoustic activity of the bees. Variables such as seasonality, weather conditions, time of day, hive configuration, and colony-specific characteristics may influence the recorded acoustic patterns and therefore contribute to the predictive behavior of the models. Although a hive-independent evaluation protocol was adopted to reduce the possibility of learning hive-specific acoustic signatures, temporal and seasonal confounding effects may still remain. For example, colony population naturally varies across seasons alongside changes in brood activity, foraging behavior, and environmental conditions, which may indirectly influence the learned representations. Similarly, differences in hive configuration, such as brood chamber-only colonies versus colonies with additional supers, may alter the acoustic environment. Finally, although the differential effect of spectral subtraction on baseline vs. modulation-based features suggests improved noise robustness for the proposed representation, a direct evaluation conditioned on labelled environmental noise events (rain, wind, speech, vehicles) could not be performed on the UrBAN dataset, which does not provide per-segment noise annotations.

Future work will therefore focus on improving the robustness and interpretability of the proposed framework under more diverse and controlled conditions. This includes season-stratified evaluation protocols, the incorporation of explicit



(a) Low (1–10 fobs). (b) Medium (11–20 fobs). (c) High (21–30 fobs).

Time frame: t_1

(d) Low. (e) Medium. (f) High.

Time frame: t_{mid}

(g) Low. (h) Medium. (i) High.

Time frame: $t_{\max}$

Fig. 8: Grad-CAM visualizations for representative 3D test samples. Each row corresponds to a different time frame (t_1, t_{mid}, $t_{\max}$), and columns correspond to colony strength levels (low, medium, high). Heatmaps show how the model's focus shifts across the modulation-frequency space over time.

environmental covariates such as temperature and humidity, normalization strategies accounting for circadian acoustic variations, and domain adaptation techniques to better separate colony strength–related acoustic signatures from environmental or seasonal effects.

V. Conclusion

In this study, we proposed the use of modulation spectrograms and tensorgrams, derived from beehive acoustic recordings, as input to deep learning neural network architectures for improved honey bee colony strength prediction. With a recent open dataset of raw audio beehive recordings, we compared the performance of two- and three-dimensional CNNs and their attention-enhanced variants, as well as convolutional recurrent CNNs. Results show that the proposed modulation tensorgrams and modulation spectrograms provide richer information for model learning, leading to improved and more generalizable prediction accuracy compared to conventional audio features. These results suggest that modulation spectral signal processing is a viable tool for automated in-the-wild beehive monitoring.

Acknowledgments

The authors acknowledge funding from NSERC via their Alliance program (ALLRP 548872-19), as well as Nectar Technologies Inc and Evan Henry for the support with the collection of the UrBAN dataset.

References

[1] Potts, S., Imperatriz-Fonseca, V., Ngo, H., Aizen, M., Biesmeijer, J., Breeze, T., Dicks, L., Garibaldi, L., Hill, R., Settele, J. & Others

Safeguarding pollinators and their values to human well-being. *Nature*. **540**, 220-229 (2016)

[2] Singh, G. & Rana, A. Honeybees and colony collapse disorder: understanding key drivers and economic implications. *Proceedings Of The Indian National Science Academy*. pp. 1-17 (2025)

[3] Mahankuda, B. & Tiwari, R. Impact of Climate Change on Honeybees and Crop Production. *Adapting To Climate Change In Agriculture-Theories And Practices: Approaches For Adapting To Climate Change In Agriculture In India*. pp. 211-224 (2024)

[4] Vanbergen, A., Initiative, I. & Others Threats to an ecosystem service: pressures on pollinators. *Frontiers In Ecology And The Environment*. **11**, 251-259 (2013)

[5] Zacepins, A., Stalidzans, E. & Meitalovs, J. Application of information technologies in precision apiculture. *Proceedings Of The 13th International Conference On Precision Agriculture*. **7** (2012)

[6] Alleri, M., Amoroso, S., Catania, P., Verde, G., Orlando, S., Ragusa, E., Sinacori, M., Vallone, M. & Vella, A. Recent developments on precision beekeeping: A systematic literature review. *Journal Of Agriculture And Food Research*. **14** pp. 100726 (2023)

[7] Chen, S., Wang, J., Lin, H., Lee, M., Liu, A., Wu, Y., Hsu, P., Yang, E. & Jiang, J. A machine learning-based multiclass classification model for bee colony anomaly identification using an IoT-based audio monitoring system with an edge computing framework. *Expert Systems With Applications*. pp. 124898 (2024)

[8] Barbisan, L., Turvani, G. & Riente, F. A Machine Learning Approach for Queen Bee Detection Through Remote Audio Sensing to Safeguard Honeybee Colonies. *IEEE Transactions On AgriFood Electronics*. (2024)

[9] Hunt, J. & Richard, F. Intracolony vibroacoustic communication in social insects. *Insectes Sociaux*. **60**, 403-417 (2013)

[10] Robles-Guerrero, A., Gómez-Jiménez, S., Saucedo-Anaya, T., López-Betancur, D., Navarro-Solís, D. & Guerrero-Méndez, C. Convolutional Neural Networks for Real Time Classification of Beehive Acoustic Patterns on Constrained Devices. *Sensors*. **24**, 6384 (2024)

[11] 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)

[12] Qandour, A., Ahmad, I., Habibi, D. & Leppard, M. Remote Beehive Monitoring Using Acoustic Signals. *Acoustics Australia*. **42**, 205 (2014)

[13] Abdollahi, M., Zhu, Y., Guimarães, H., Coallier, N., Maucourt, S., Giovenazzo, P. & Falk, T. On the Prediction of Varroa Mite Infestations in Honeybee Colonies via Acoustic Monitoring. *IEEE Sensors Journal*. pp. 1-1 (2026)

[14] Janetzky, P., Schaller, M., Krause, A. & Hotho, A. Swarming Detection in Smart Beehives Using Auto Encoders for Audio Data. *2023 30th International Conference On Systems, Signals And Image Processing (IWSSIP)*. pp. 1-5 (2023)

[15] Zhu, Y., Abdollahi, M., Maucourt, S., Coallier, N., Guimarães, H., Giovenazzo, P. & Falk, T. Early prediction of honeybee hive winter survivability using multi-modal sensor data. *IEEE International Workshop On Metrology For Agriculture And Forestry*. pp. - (2023)

[16] Zhu, Y., Abdollahi, M., Maucourt, S., Coallier, N., Guimarães, H., Giovenazzo, P. & Falk, T. MSPB: a longitudinal multi-sensor dataset with phenotypic trait measurements from honey bees. *Scientific Data*. **11**, 860 (2024)

[17] Abdollahi, M., Henry, E., Giovenazzo, P. & Falk, T. The importance of context awareness in acoustics-based automated beehive monitoring. *Applied Sciences*. **13**, 195 (2022)

[18] Mattila, H., Kernen, H., Otis, G., Nguyen, L., Pham, H., Knight, O. & Phan, N. Giant hornet (Vespa soror) attacks trigger frenetic antipredator signalling in honeybee (Apis cerana) colonies. *Royal Society Open Science*. **8** (2021)

[19] Cejrowski, T., Szymański, J. & Logofătu, D. Buzz-based recognition of the honeybee colony circadian rhythm. *Computers And Electronics In Agriculture*. **175** pp. 105586 (2020)

[20] Ruvinga, S., Hunter, G., Nebel, J. & Duran, O. Prediction of honeybee swarms using audio signals and convolutional neural networks. *Workshops At 18th International Conference On Intelligent Environments (IE2022)*. pp. 146-154 (2022)

[21] Kulyukin, V., Mukherjee, S. & Amlathe, P. Toward audio beehive monitoring: Deep learning vs. standard machine learning in classifying beehive audio samples. *Applied Sciences*. **8**, 1573 (2018)

[22] Phan, T., 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*. **27**, 5873-5892 (2023)

[23] 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)

[24] Nolasco, I., Terenzi, A., Cecchi, S., Orcioni, S., Bear, H. & Benetos, E. Audio-based identification of beehive states. *ICASSP 2019-2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 8256-8260 (2019)

[25] Campell, C., Parry, R., Tashakkori, R., Somer, A., Richardson, L. & Simons-Rudolph, A. Honey Bee Piping Detection Utilizing Convolutional Neural Networks. *IEEE Sensors Journal*. (2025)

[26] Kim, J., Oh, J. & Heo, T. Acoustic Scene Classification and Visualization of Beehive Sounds Using Machine Learning Algorithms and Grad-CAM. *Mathematical Problems In Engineering*. **2021** (2021)

[27] Abdollahi, M., Zhu, Y., Guimarães, H., Coallier, N., Maucourt, S., Giovenazzo, P. & Falk, T. Audio Modulation Spectral Features for Improved Honeybee Colony Population Prediction. *IEEE Sensors Journal*. pp. 1-1 (2025)

[28] Bricout, A., Leleux, P., Acco, P., Escriba, C., Fourniols, J., Soto-Romero, G. & Floquet, R. Bee together: Joining bee audio datasets for hive extrapolation in AI-based monitoring. *Sensors*. **24**, 6067 (2024)

[29] Tiwari, A., Cassani, R., Kshirsagar, S., Tobon, D., Zhu, Y. & Falk, T. Modulation spectral signal representation for quality measurement and enhancement of wearable device data: A technical note. *Sensors*. **22**, 4579 (2022)

[30] Gulli, A. & Pal, S. Deep learning with Keras. (Packt Publishing Ltd,2017)

[31] Abdollahi, M., Zhu, Y., Guimarães, H., Coallier, N., Maucourt, S., Giovenazzo, P. & Falk, T. UrBAN: Urban Beehive Acoustics and PheNotyping Dataset. *Scientific Data*. **12**, 536 (2025)

[32] Chabert, S., Requier, F., Chadoeuf, J., Guilbaud, L., Morison, N. & Vaissiere, B. Rapid measurement of the adult worker population size in honey bees. *Ecological Indicators*. **122** pp. 107313 (2021)

[33] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal Of The American Statistical Association*. **32**, 675-701 (1937)

[34] Nemenyi, P. Distribution-free multiple comparisons.. (Princeton University,1963)

[35] Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 618-626 (2017)