S.T. Yau High School Science Award

Research Report

The Team

Name of team member: Andrew Liang
School: The Harker School
City, Country: San Jose, USA

Name of supervising teacher: Ms. Anuradha Datar
Job Title: Teacher, Computer Science Department
School/Institution: The Harker School
City, Country: San Jose, USA

Title of Research Report

An Intelligent Bee Health Assessment System with Cross-Attention Multimodal Integration of Visual and Audio Data

Date

08/22/24

An Intelligent Bee Health Assessment System with Cross-Attention Multimodal Integration of Visual and Audio Data

Andrew Liang

Abstract

Honeybees play a crucial role in pollinating approximately one-third of the world's food supply. However, bee colonies have suffered a nearly 40\% decline over the past decade due to threats such as parasites like the varroa mite. Traditional beehive monitoring methods, including human inspections, are often subjective, disruptive, and time-consuming. Machine learning models have been applied to evaluate beehive health, but previous studies have relied on single-source data, such as bee images or sounds, and lacked complete solutions. This study introduces an innovative system for bee object detection and health evaluation, leveraging both visual and audio signals to analyze various bee behaviors. The system uses a Cross-Attention based Multimodal Neural Network (CAMNN) to adaptively highlight key features from each signal, achieving 80.7\% accuracy and significantly outperforming existing single-signal models across four health categories. Moreover, it demonstrates strong prediction robustness, maintaining an F1-score above 70\% across all four evaluated health conditions. The CAMNN-based beehive monitoring and health assessment system can track bee activities and deliver early alerts for potential hive collapse.

Keywords: Cross-Attention Multimodal Neural Network (CAMNN), bee object detection, bee health assessment, computer vision, signal processing, electronic beehive monitoring

Acknowledgement

I would like to extend my sincere gratitude to the Santa Clara Valley Beekeepers Guild, Gilroy Beekeepers Association, Small Bee, and Randy Oliver, a renowned researcher and author on beekeeping in California, USA. Their support in data collection and their insightful feedback have been invaluable to this research.

I would like to extend my sincere gratitude to my mentor, Ms. Anuradha Datar at Harker School, for her guidance, encouragement, and support throughout this research.

Commitments on Academic Honesty and Integrity

We hereby declare that we

1. are fully committed to the principle of honesty, integrity and fair play throughout the competition.
2. actually perform the research work ourselves and thus truly understand the content of the work.
3. observe the common standard of academic integrity adopted by most journals and degree theses.
4. have declared all the assistance and contribution we have received from any personnel, agency, institution, etc. for the research work.
5. undertake to avoid getting in touch with assessment panel members in a way that may lead to direct or indirect conflict of interest.
6. undertake to avoid any interaction with assessment panel members that would undermine the neutrality of the panel member and fairness of the assessment process.
7. observe the safety regulations of the laboratory(ies) where we conduct the experiment(s), if applicable.
8. observe all rules and regulations of the competition.
9. agree that the decision of YHSA is final in all matters related to the competition.

We understand and agree that failure to honour the above commitments may lead to disqualification from the competition and/or removal of reward, if applicable; that any unethical deeds, if found, will be disclosed to the school principal of team member(s) and relevant parties if deemed necessary; and that the decision of YHSA is final and no appeal will be accepted.

(Signatures of full team below)

X Andrew Liang

Name of team member:

X. Ms. Anuradha Datar

Name of supervising teacher:

Table of Contents

An Intelligent Bee Health Assessment System with Cross-Attention Multimodal Integration of Visual and Audio Data

Andrew Liang

***Abstract*— Honeybees play a crucial role in pollinating approximately one-third of the world's food supply. However, bee colonies have suffered a nearly 40% decline over the past decade due to threats such as parasites like the varroa mite. Traditional beehive monitoring methods, including human inspections, are often subjective, disruptive, and time-consuming. Machine learning models have been applied to evaluate beehive health, but previous studies have relied on single-source data, such as bee images or sounds, and lacked complete solutions. This study introduces an innovative system for bee object detection and health evaluation, leveraging both visual and audio signals to analyze various bee behaviors. The system uses a Cross-Attention based Multimodal Neural Network (CAMNN) to adaptively highlight key features from each signal, achieving 80.7% accuracy and significantly outperforming existing single-signal models across four health categories. Moreover, it demonstrates strong prediction robustness, maintaining an F1-score above 70% across all four evaluated health conditions. The CAMNN-based beehive monitoring and health assessment system can track bee activities and deliver early alerts for potential hive collapse.**

***Index Terms*— Cross-Attention Multimodal Neural Network (CAMNN), bee object detection, bee health assessment, computer vision, signal processing, electronic beehive monitoring**

I. Introduction

HONEYBEES are essential to global agriculture, contributing approximately $500 billion annually by pollinating crops. In the United States, their pollination services were valued at $15 to $20 billion in 2020 [1]. Alarmingly, U.S. bee populations have dropped by nearly 40% over the past decade [2], posing a significant threat to agricultural productivity and food supply.

Andrew Liang is with The Harker School, San Jose, CA USA; alcago2020@gmail.com

Honey bee colonies face numerous challenges worldwide, including parasites like varroa mites, missing queens, and colony collapse. Varroa mites, notorious parasites, weaken bees by feeding on their bodily fluids and spreading harmful viruses [3]. The absence of a queen disrupts hive productivity, leading to decreased hive population and potential collapse [4]. Swarming results in a substantial portion of the colony departing with the queen to form a new hive, leaving the original hive weak and vulnerable [5]. Detecting beehive stress early is crucial for preserving pollination services and maintaining healthy ecosystems.

Beekeepers can identify potential threats and take appropriate action by tracking bee populations and observing various bee behaviors. However, traditional methods, such as human inspection and bee sampling, are often subjective, intrusive, and labor-intensive. Moreover, these approaches may miss subtle changes in bee behavior, leading to delayed detection of health issues. Recently, alternative methods like sensor technology have provided a non-disruptive way to continuously monitor bee activities. However, these advancements primarily offer basic monitoring functions without providing in-depth insights into hive health.

Recent advancements in machine learning have driven innovation in beekeeping. Computer vision techniques can analyze images of individual bees at the hive entrance to detect changes in behavior, wing symmetry, and body morphology, which can serve as early indicators of stress, disease, or hive decline. Additionally, audio analysis of beehive recordings, especially looking at the frequency and intensity of bee sounds, offers valuable insights into colony activities, such as queen presence, and potential swarming events. Despite these advancements, there is still a lack of unified solutions that integrate bee object detection with health assessments. Furthermore, the potential benefits of combining diverse data sources for

enhanced hive health evaluation have not been investigated. Moreover, there is a lack of high-quality datasets with bee images and audio recordings, and none that pair images with audio for comprehensive analysis. To address these challenges, this study introduces a two-step system that not only identifies bees in images and audio but also assesses their health. By using convolutional blocks, the system extracts visual and audio features, ensuring that only relevant data is used for evaluation. Inspired by the cross-attention mechanism [6] that is widely adopted in natural language processing to capture relationships between different modalities or sequences, The Cross-Attention based Multimodal Neural Network (CAMNN) is designed to integrate bee visual and audio signals and classify the hive health into different categories. By allowing the image sequence (e.g., query) to attend to the corresponding audio sequence (e.g., key-value pair) through attention scores, the model can focus on the most relevant information from each data modality and improve the performance.

To summarize, the main contributions of this study are:

1) This study is the first to integrate visual and audio signals for analyzing bee behavior more effectively.
2) A two-step framework is presented that not only identifies individual bees but also assesses their health using only relevant data.
3) The CAMNN improves bee health evaluation accuracy by enabling image features to query corresponding relevant audio features.
4) Two high-quality datasets for bee object detection are made publicly available. The remaining two datasets for health assessment will be released after patent approval.

II. Related Work

This section explores four areas that are relevant to the study: the relationship between visual and acoustic cues and bee health, computer vision and audio signal processing in bee research, cross-attention mechanism, and advancements in beehive monitoring systems. These insights guide the development of a more effective and integrated beehive management system.

A. Bee Colony Health Assessment through Visual and Acoustic Cues

The overall health of a bee colony can often be assessed by analyzing the physical appearance of individual bees [7]. Healthy bees are typically lively with a shiny and smooth exoskeleton. Visual signs such as damaged wings, varroa mites attached to their bodies, unusual spots, or discoloration can indicate diseases like the deformed wing virus [8], infestations by parasites [9], fungal infections, or chemical exposure [10]. Physical traits such as size and symmetry provide insights into their nutritional health and development [11]. Additionally, observing pollen loads on foragers provides information about foraging success and pollen diversity, which are crucial indicators of overall colony health [12].

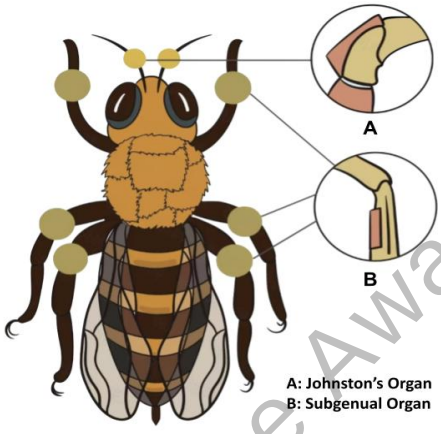

Fig. 1. The primary receptors for detecting vibroacoustic signals

Bees produce sounds not only with their wings but also by vibrating their thoracic muscles. Contrary to earlier beliefs that bees were deaf to airborne sounds, recent research shows they can detect air particle movements through vibroacoustic reception using Johnston's organs in their antennae and subgenual organs in their legs, as illustrated in Figure 1 [13]–[18]. Experienced beekeepers can assess the state of a colony by noting variations in bee sounds [19]. Burgess et al. [20] use acoustic measurements to detect early stages of varroa mite infestation, indicated by altered acoustic behavior of the beehive following relevant stress exposure. Other researche also indicates that honey bee sounds provide insights into colony conditions such as swarms, and queen presence, each associated with specific frequency patterns. Table I outlines these bee sounds, their frequency ranges, sources, and implications based on the research in [21]–[24].

B. Computer Vision and Audio Signal Processing in Bee Research

Images and videos of bees provide valuable insights into colony activity and health [25]. Researchers often collect data from beehive entrances, utilizing various machine learning models to identify bees within complex environments. These models also analyze individual bee images to detect hive issues

TABLE I
FREQUENCY RANGES AND SIGNIFICANCE OF BEE SOUND

Signals	Frequency (Hz)	Producer	Implications
Piping	330 $\sim$ 430	Worker bees	Normal, stress-free activity
Piping	100 $\sim$ 500	Worker bees	Imminent or ongoing swarm
Hissing	300 $\sim$ 3600	Worker bees	Halts behaviors such as forager dancing and hive departures
Piping	200 $\sim$ 550	Young queens	Young queen ready to emerge
Tooting	400 $\sim$ 500	Young queens	Young queen has emerged
Quacking	$\sim$ 350	Mature queens	Multiple queens present when the first young queen emerges

such as missing queens, hive robberies, ant and parasite infestations [26], [27]. For example, Ratnayake et al. [28] develop a hybrid algorithm that combines image-based tracking with background subtraction and deep learning, achieving an accuracy rate of 86.6%. Liu et al. [29] collect 2,000 high-resolution images of bees with varroa mites and proposed a convolutional neural network (CNN) to detect varroa destructor, with an accuracy of 96.85%. Other studies [30]–[33] employ machine learning models such as SVM, Inceptions, MobileNet, on a dataset comprising 5,172 individual bee images captured at hive entrances. These models classify beehive health into six categories: Healthy, Missing Queen, Few Varroa Mites and Hive Beetles, Varroa Mites and Small Hive Beetles, Ant Problems, and Hive Robbery, all achieving accuracy rates exceeding 90%.

Researchers explore various signal processing techniques to detect and analyze bee sounds, which can indicate colony conditions such as swarming, missing queens, and pest infestations [25]. Kim et al. [34] use VGG-13 combined with MFCC audio features, achieving 91.93% accuracy in detecting bee sounds. Terenzi et al. [35] use CNN combined with multiple audio features, including Short-time Fourier transform (STFT), Mel spectrograms, HHT, discrete wavelet transform (DWT), and Continuous Wavelet Transform (CWT), on the Nu-hive data [36] identifying missing queens with 78.58% accuracy using Mel spectrograms. The Open Source BeeHive (OSBH) project [37] collect audio recordings from personal beehives as part of a citizen science initiative, with participants also contributing metadata. The OSBH project's diverse recordings, including variations in recording devices, environments, and microphone placements, provide valuable data for real-world evaluation. Zgank et al. [38] apply Mel-frequency cepstral coefficients (MFCC) and hidden Markov on the OSBH data to detect swarm behavior, achieving 80.89% accuracy.

C. Cross-Attention Multimodal Models in Other Fields

The cross-attention mechanism, originally developed for natural language processing, is increasingly being used to merge embedding sequences from different modalities. Recently, it has been applied across various domains, including computer vision and audio processing, to enhance the performance of multimodal tasks by effectively integrating information from diverse data sources. Qian et al. [39] introduce an audio-visual Cross-Modal Attentive Fusion (CMAF) mechanism for robotic speaker tracking, which leverages self-attention for temporal alignment and cross-attention for inter-modal alignment, achieving a 5.82% improvement in accuracy. Chen et al. [40] develop a token fusion module based on cross attention, which uses a single token for each branch as a query to exchange information with other branches, for image classification. The model achieves a 2% improvement over DeiT on the ImageNet1K dataset. Additionally, Khattar et al. [41] use a cross-attention mechanism to integrate text and image data, outperforming unimodal and bilinear models by 5.91% to 6.31% in multimodal disaster classification tasks. These advancements underscore the potential of cross-attention in combining visual and audio signals, making it a promising technique for applications in beehive health monitoring.

D. Beehive Monitoring System

Researchers continue to advance various IoT-based systems to monitor and manage bee colonies. These systems employ remote sensors to provide real-time data on hive conditions, significantly enhancing beekeeping practices and colony health management. Mrozek et al. [42] develop an embedded system for monitoring varroa mites, featuring a camera positioned in front of the beehive, powered by a Raspberry Pi 4 and a Coral TPU accelerator. Tashakkori et al. [43] implement a video surveillance system with cameras positioned in front of or above hives, aimed at the entrance. These cameras rely on Frequency Hopping Spread Spectrum (FHSS) technology for

digital signal transmission and include microphones to capture environmental sounds, such as those linked to swarming. Additionally, the Bee Health Guru [44] utilizes smartphones to monitor hive sounds, which artificial intelligence then processes to assess hive conditions, including queen loss, varroa mite infestations, or small hive beetles.

III. Methods

The framework for bee object detection and health assessment, as illustrated in Figure 2, is developed through the following steps in the study:

1) Data Acquisition and Annotation: Collect and annotate images and audio clips of bees for object detection and health assessment.
2) Data Processing: Extract features from the audio clips and perform data augmentation.
3) Bee Object Detection: Utilize object detection algorithms to identify bees in images and audio clips, ensuring that only data containing bees is used for health assessment.
4) Bee Health Assessment: Evaluate bee health using visual and audio signals with cross-attention mechanism to detect signs of beehive stress.

A. Data Acquisition

The data acquisition aims to collect both visual and acoustic signals from bees. Videos were recorded at the entrances of thirty beehives located in apiaries in California (approximate coordinates: 37.3387° N, 121.8853° W) from October 2022 to June 2023. This approach ensures minimal disruption to the bees' activities while capturing comprehensive data. The beehives house colonies of Apis mellifera with each colony consisting of approximately 60,000 bees.

To capture images, an Arducam IMX519 Raspberry Pi camera was positioned above beehive entrances. The camera had a back-illuminated stacked sensor with a pixel size of $1.22\,\mu\text{m} \times 1.22\,\mu\text{m}$ and an autofocus lens with an aperture of $f/1.75$, enabling it to capture focused and high-resolution images, as well as record videos in 720p at 60 frames per second. The camera was placed to ensure that bees entering or leaving the hive would be observed.

In addition to visual data, audio was captured using a high-quality PoP voice professional microphone positioned near the beehive entrance. With a high audio sensitivity of 30 dB and a frequency response of 20 Hz to 20 kHz, the microphone ensured that sounds of various frequencies produced by the bees and their surroundings were captured. Noise cancellation and a windscreen feature minimized background noise, further enhancing the data quality.

Next, the videos are divided into frames and audio clips for annotation. Frames are extracted at ten-second intervals, and the audio clips are standardized to ten seconds in length. If the video duration is not a multiple of ten seconds, any remaining part is discarded. The frames and corresponding audio clips are then paired. To maintain a diverse and representative sample set, only a random selection of these paired frames and audio clips is used for further study.

B. Data Annotation

1) Annotation for Bee Image Object Detection: The image annotation is conducted using the Label Studio platform. For each image, individual bees are manually annotated by drawing bounding boxes (BBox) around their bodies and wings, as illustrated in Figure 3. The output label files are generated in the YOLO format, which records the BBox coordinates for each annotated bee. The label file contains the object ID, X-axis center, Y-axis center, BBox width, and height. All values are normalized to the image size, ranging from 0 to 1. In cases when multiple bees are present in an image, each bee is represented by a separate line in the label file.

To ensure consistency and accuracy, only images meeting the following conditions are included in the final dataset:

- Bee visibility: Each image must show at least 50% of the bee's body.
- Image quality: Only clear bee images are included to maintain high data quality.

During the labeling process, bees are labeled as completely as possible. Additionally, two annotators randomly cross-check the annotations, ensuring that the datasets contain detailed and reliable data.

2) Annotation for Bee Audio Object Detection: During audio annotation, the bee sounds and other background noises, such as those from birds, airplanes, and cars, are flagged. Audio clips are labeled as bee-related if bee sounds dominate the recording. Each recording is carefully listened to twice to ensure the accuracy.

3) Annotation for Bee Health Classification: Bee images and audio clips are annotated with natural beehive health statuses verified by experienced beekeepers. Each image and audio clip is assigned one of four labels: healthy bees, beehives with swarms, missing queens, or varroa mites.

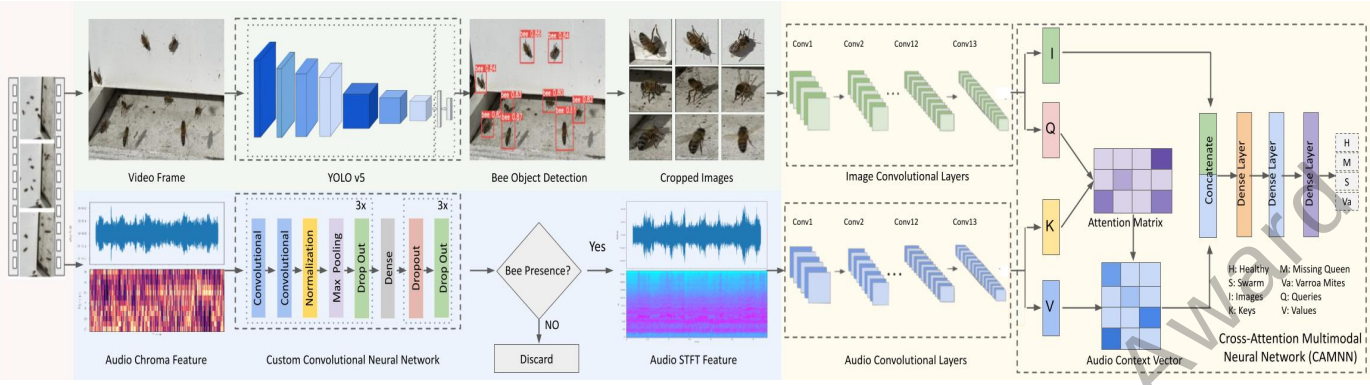

Fig. 2. The proposed framework for bee object detection and health assessment: The video is split into images and audio clips. Bees are identified and cropped from the images, and audio clips containing bee sounds are identified. Visual and audio features are then extracted from these bee-containing data using convolutional layers. Finally, these features are integrated using the Cross-Attention-based Multimodal Neural Network (CAMNN) for bee health assessment.

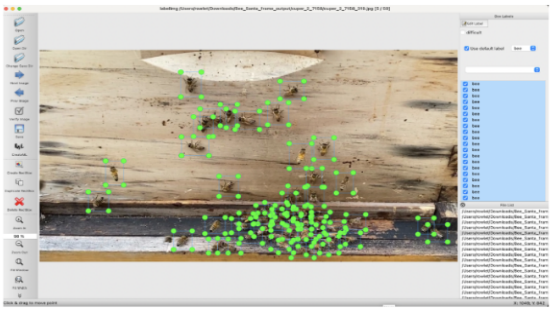

Fig. 3. Bee image object detection annotation in the Label Studio

TABLE II
FOUR DATASETS FOR BEE DETECTION AND HEALTH ASSESSMENT USING IMAGES AND AUDIO CLIPS

Data Type	Purpose	Label	Count
Images	Bee Object Detection	Bee/No Bee	1,524
Audio Clips			2,840
Images & Audio Clips	Health Assessment	Healthy	1,960
		Swarm	1,896
		Varroa Mites	1,722
		Missing Queen	1,886

Finally, this study produces four distinct datasets: one image and one audio dataset for bee object detection, and one image and one audio dataset for bee health evaluation. These datasets capture various aspects of bee behavior and health, as summarized in Table II.

C. Audio Feature Extraction

After loading the audio samples, four features are extracted to analyze the acoustic characteristics for bee identification and health classification. These features include the Mel Spectrogram [45], Mel-Frequency Cepstral Coefficients (MFCC) [46],

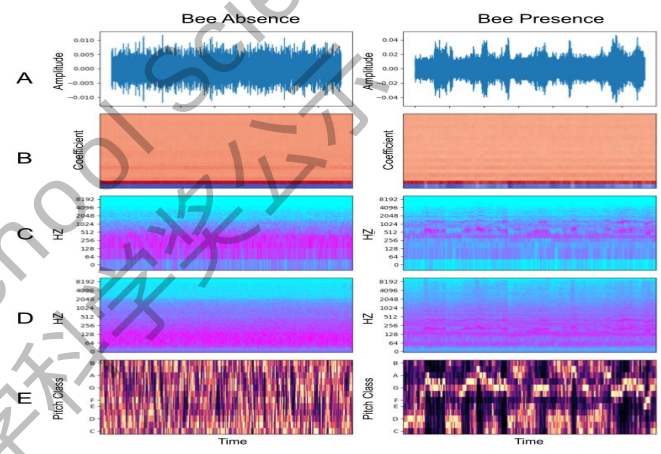

Fig. 4. The audio features show distinct patterns when bees are absent (left) vs. present (right). The top-to-bottom representation includes (A) original audio wave, (B) Mel Spectrogram, (C) MFCC, (D) STFT, and (E) Chromagram.

Short-Time Fourier Transform (STFT) [47], and Chromagram [48]. They effectively capture the temporal dynamics, spectral content, timbral texture, and harmonic properties of bee sounds. After converting the audio clips into these features, distinct patterns are revealed in various scenarios. Figure 4 visually displays these audio features, highlighting differences in the presence or absence of bees.

D. Data Augmentation

Data augmentation is commonly used in machine learning to artificially increase the variability of a training dataset by applying various transformations to the existing data. This process helps improve model generalization and robustness.

1) Visual Data Augmentation: The image datasets are augmented by applying independent transformations, including rotating from 0 to 360 degrees, zooming by up to 20%, horizontal and vertical shifting up to 20% of the total width and

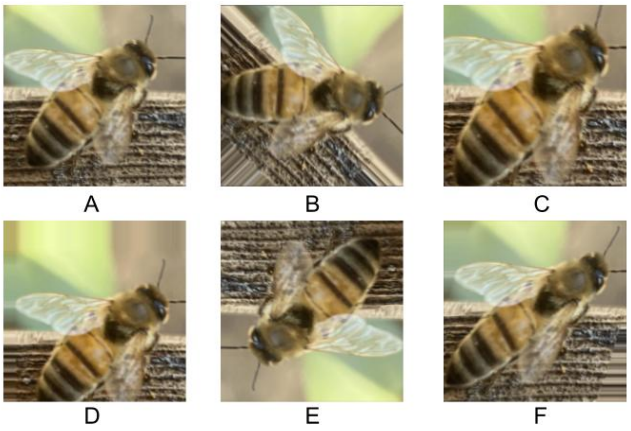

Fig. 5. Image data augmentation in training dataset to improve model robustness. (A) Original, (B) Rotation, (C) Zoom, (D) Shift, (E) Flip, (F) Shear

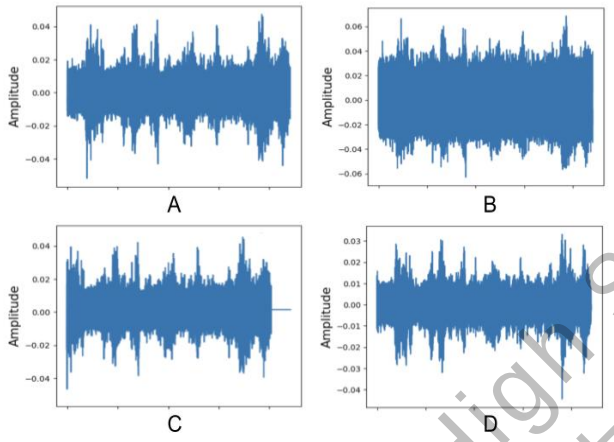

Fig. 6. Audio data augmentation in training dataset to improve model robustness. (A) Original, (B) Random Noise, (C) Shift, (D) Pitch

height, random horizontal and vertical flipping, and shearing up to 20% distortion, as illustrated in Figure 5. Subsequently, all the RGB images are resized to 224x224 pixels.

2) Audio Data Augmentation: Various methods are used to augment the audio data. White noise is added by setting its amplitude to a random value between 0 and 0.05. The audio signal is also randomly shifted within a range of -5 to 5 milliseconds to simulate real-world variations. Furthermore, random pitch shifts are applied within a range of -2 to 2. The data augmentation is depicted in Figure 6.

E. Bee Object Detection

1) Bee Image Object Detection: YOLO [49] is a model designed for fast and accurate object detection. Several YOLOv5 model configurations, such as YOLOv5s6, YOLOv5m6, and YOLOv5l6, are evaluated for their ability to localize bees in images. The YOLOv5m6 model, comprising 276 layers and 35,248,920 trainable parameters, demonstrates the best performance. After detecting the bees, the model generates bounding boxes used to crop the bees for further analysis. Figure 2 illustrates the identification and cropping of nine bees, each assigned a probability score.

2) Bee Audio Object Detection: The custom convolutional neural network (CNN) models, each utilizing a different audio feature, are developed to identify bees in audio clips. The model begins with two convolutional layers, each with 64 filters and an 8-unit kernel, to capture bee sound frequencies. Batch normalization is then implemented to expedite model training, followed by a max-pooling layer to reduce data dimensions and a dropout layer with a rate of 25% to reduce overfitting. This pattern is repeated twice: first with two convolutional layers of size 128, then with another two convolutional layers of size 256, each followed by batch normalization, max pooling, and a dropout layer. NNext, the data are flattened and passed through three dense layers with 32, 64, and 128 neurons, respectively, each followed by a 25% dropout layer. The model ends with a 2-unit dense layer using softmax activation to classify sounds as either the presence or absence of bees, as illustrated in Figure 2. The model contains 7,739,236 parameters, among which 7,738,340 are trainable.

F. Bee Health Assessment

1) Visual and Audio Feature Extraction: To identify the most effective method for extracting visual features from images, four model structures are tested: VGG16 [50], MobileNet v2 [51], Inception v3 [52], and the custom CNN previously employed in audio object detection. Similarly, for extracting features from audio clips, three models are assessed: VGG16, LSTM [53], and the custom CNN used for audio object detection, each combined with the STFT audio feature. VGG16, a pre-trained deep CNN, is selected for both visual and audio feature extraction due to its robust performance.

The image feature maps, shown in Figure 7, are generated from the convolutional layers of the VGG16 model. These maps capture progressively complex visual patterns, from basic edges in the early layers to intricate shapes in the deeper layers. These progression enhances the network's ability to effectively identify visual content.

2) CAMNN for Bee Health Assessment: Previous studies on bee health assessment have primarily focused on either images or audio alone. To address the limitations of these methods, a cross-attention mechanism [6] is applied to combine visual and auditory information. This mechanism enables the visual

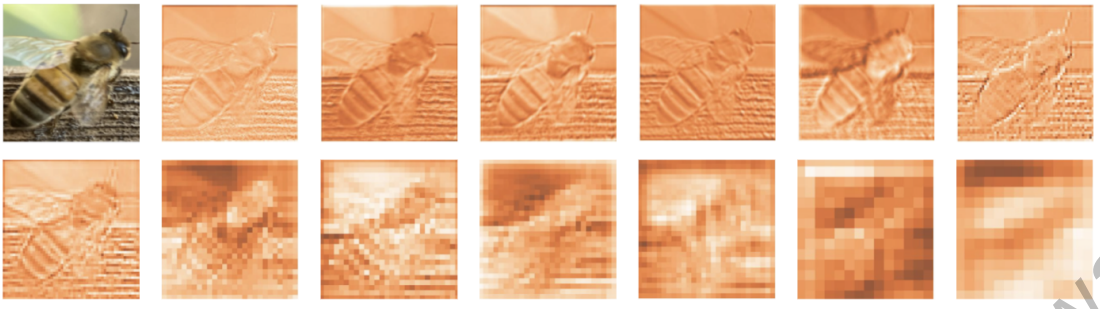

Fig. 7. Visualization of bee image feature maps, highlighting key regions of an input image across the convolutional layers

signals to dynamically retrieve corresponding important details from the audio features by leveraging attention scores, improving the model's ability to capture complex patterns related to bee health.

In the study, the image features ($\mathbf{I}$) and auditory features ($\mathbf{A}$) are first transformed into queries, keys, and values. The queries ($\mathbf{Q}$) are generated from the image features, while the keys ($\mathbf{K}$) and values ($\mathbf{V}$) are generated from the auditory features:

$$\mathbf{Q} = \mathbf{I}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{A}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{A}\mathbf{W}_V$$

Here, $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable weight matrices that project the image and auditory features into a shared space for calculating attention scores. The attention scores ($\mathbf{S}$) are computed as the dot product of $\mathbf{Q}$ and $\mathbf{K}^\top$, scaled by $\sqrt{d_k}$, and normalized using the softmax function:

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}, \quad \mathbf{A}_{\text{softmax}} = \text{softmax}(\mathbf{S})$$

Here, d_k is the dimension of the keys in the attention mechanism. These scores are then used to weight the value matrix ($\mathbf{V}$) to generate a context vector ($\mathbf{C}$):

$$\mathbf{C} = \mathbf{A}_{\text{softmax}}\mathbf{V}$$

The context vector is concatenated with the original visual embeddings to create a fused representation ($\mathbf{F}$), which leverages both visual and audio features. This fused representation is then passed through dense layers with ReLU activation functions and dropout for regularization. The final output layer uses a softmax function to predict bee health conditions, as detailed in Algorithm 1.

During model training, the weights are optimized to minimize the loss function, which has two main components: the reconstruction loss for both image and sound modalities. The loss function integrates these components, assigning different levels of importance to each, as described below.

Algorithm 1 Algorithm for the Cross-Attention MultiModal Model

function BEEHEALTHMODEL(image_input, audio_input)
    feature_image ← VGG16(image_input)
    feature_audio ← VGG16(audio_input)
    query ← Dense(d_k)(feature_image)
    key ← Dense(d_k)(feature_audio)
    value ← Dense(d_v)(feature_audio)
    attention_scores ← softmax(dot(query, transpose(key)) / $\sqrt{d_k}$)
    context_vector ← dot(attention_scores, value)
    feature_concatenated ← concatenate(feature_image, context_vector)
    fc_layer1 ← Dense(32, *relu*)(feature_concatenated)
    fc_layer1_dropout ← Dropout(0.5)(fc_layer1)
    fc_layer2 ← Dense(16, *relu*)(fc_layer1_dropout)
    fc_layer2_dropout ← Dropout(0.5)(fc_layer2)
    output ← Dense(4, *softmax*)(fc_layer2_dropout)
    return output
end function

$$L = \lambda_{\text{image}} L_{\text{image}}(y, \hat{y}_{\text{image}}) + \lambda_{\text{sound}} L_{\text{sound}}(y, \hat{y}_{\text{sound}})$$

$$L_{\text{image}}(y, \hat{y}_{\text{image}}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j} \log(\hat{y}_{\text{image},i,j})$$

$$L_{\text{sound}}(y, \hat{y}_{\text{sound}}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j} \log(\hat{y}_{\text{sound},i,j})$$

Where:

L : The loss function

y : The true label for the j-th record in class i.

$\hat{y}$: The predicted label for the j-th record in class i

λ : The importance for the j-th record in class i

N : The number of classes

M : The number of records in each class

TABLE III
HYPERPARAMETERS FOR BEE OBJECT DETECTION AND BEE HEALTH ASSESSMENT

Hyperparameters	Bee Object Detection		Bee Health Assessment
	Image	Audio	
Epochs	50	50	100
Batch Size	64	128	64
Early Stop	20	20	20
Learning Rate	0.01	0.001	0.0001
Optimizer	SGD*	Adam	Adam
Momentum	0.9	NA	NA

*Stochastic Gradient Descent

IV. EXPERIMENTAL RESULTS

The evaluation employs five-fold cross-validation, where the dataset is randomly divided into five equal segments. In each iteration, one segment is used for testing while the other four are used for training. This process is repeated five times, each with a different segment used for testing. The model's performance is then evaluated using several metrics, such as accuracy, precision, recall, and the F1-score. The model hyperparameters are fine-tuned to optimize the categorical cross-entropy loss function. The hyperparameters that generate the best model performance are listed in Table III. All models are developed on an NVIDIA Tesla T4 GPU, which comes with 16GB of GDDR6 memory and 2,560 CUDA cores.

A. Bee Image Object Detection

In the study, YOLOv5m6 is used to locate and crop bees within pictures. The model demonstrates high accuracy, achieving 94.8% precision, 95.9% recall, 98.6% mean average precision (mAP@50) and 70.4% mean average precision (mAP@50-95). The rate mAP@50 is assessed by comparing how closely the locations identified by the model match the real locations of the bees when at least half of the area overlaps. The accuracy is further evaluated under varying conditions, where the overlap between the predicted and actual locations of bees ranges from 50% to 95%. This thorough assessment demonstrates the model's ability to accurately detect bees. The mAP@X is calculated by:

$$\text{mAP@X} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{1}$$

where N is the number of classes and AP_i is the average precision at an intersection over Union (IoU) threshold of X for each class.

TABLE IV
PERFORMANCE METRICS OF FOUR AUDIO FEATURE EXTRACTION TECHNIQUES FOR BEE SOUND DETECTION

Audio Features	Accuracy	Precision	Recall	F1-Score
Chromagram	**95.1%**	**95.1%**	**95.1%**	**95.1%**
STFT	89.4%	89.5%	89.4%	89.4%
MFCC	82.4%	82.8%	82.4%	82.3%
Mel Spectrogram	76.4%	83.4%	76.4%	75.1%

TABLE V
F1-SCORES FOR BEE SOUND DETECTION IN FOUR AUDIO FEATURE EXTRACTION TECHNIQUES

Audio Features	Bee	No_Bee
Chromagram	**91.2%**	**96.6%**
STFT	89.3%	89.6%
MFCC	81.3%	83.3%
Mel Spectrogram	80.8%	69.4%

B. Bee Audio Object Detection

The custom CNN is applied to detect bee sounds in audio recordings using four distinct audio features. This lead to varying levels of accuracy, as shown in Table IV. Mel Spectrogram captures broad spectral properties but may overlook fine details, resulting in 76.4% accuracy and a 75.1% F1-score. MFCC excels in capturing the timbral aspects of bee sounds, achieving 82.4% accuracy and F1-score. The STFT is effective in analyzing short-term changes in both frequency and time, making it well-suited for detecting rapid variations in buzzing. It achieves an accuracy and F1-score of 89.4%. The Chromagram excels in identifying harmonic patterns and performs well in scenarios like the bees' waggle dance, where harmonic elements are prominent. It achieves the highest accuracy and F1-score, both at 95.1%. The effectiveness of each method depends on its ability to capture the distinct biological characteristics of bee sounds linked to different behaviors.

The models perform consistently in detecting both the presence and absence of bee sounds, except for the one using Mel Spectrogram. The model based on Chromagram achieves similar F1-scores of 91.2% and 96.6% for both bee and non-bee sounds, as shown in Table V. This indicates that the models are equally proficient at recognizing bee and non-bee sounds. Such a balance is crucial when it is equally important to avoid mistakenly classifying non-bee sounds as bee sounds and missing actual bee sounds.

TABLE VI
F1-SCORES FOR EACH BEE HEALTH CONDITION IN DIFFERENT MODEL CONFIGURATIONS

Models	Healthy	Varroa Mites	Swarm	Missing Queen
CAMNN	**89.8%**	**71.9%**	**78.9%**	**80.9%**
w/o cross-attn.	86.3%	70.7%	73.1%	75.4%
w/o image	76.9%	56.8%	72.7%	72.1%
w/o audio	68.6%	68.2%	65.9%	45.1%

TABLE VII
ABLATION STUDY RESULTS

Models	Accuracy	Precision	Recall	F1-score
CAMNN	**80.7%**	**82.5%**	**80.7%**	**80.8%**
w/o cross-attn.	77.3%	78.0%	77.3%	77.4%
w/o image	70.5%	72.5%	70.5%	71.0%
w/o audio	63.1%	64.2%	63.1%	63.2%

C. Multimodal Bee Health Assessment

1) CAMNN Model Performance: A cross-attention based multimodal neural network (CAMNN) has been developed to combine visual and acoustic signals for more effective assessment of bee health, achieving an accuracy of 80.7%. The CAMNN model not only improves overall accuracy but also significantly boosts the classification of specific bee health conditions, as demonstrated in Table VI. For example, the F1-score for detecting bees affected by a missing queen increases from 45.1% without audio to 80.9% with the CAMNN model. Similar improvements are observed in the other three bee health categories. As a result, the model consistently achieves high performance, maintaining F1-scores above 70% for all bee health states.

2) Ablation Study: An ablation study is conducted to evaluate the effectiveness of CAMNN by individually removing cross attention mechanism, visual and audio signals. Table VII demonstrates that the model's performance decreases by 4.2%, 12.7%, and 21.8% when the cross-attention mechanism, audio signals, and visual signals are removed, respectively. This ablation study demonstrates the interdependence of these components, confirming that CAMNN's strength lies in the complementary nature of visual and audio signals, enhanced by the cross-attention mechanism.

An interesting observation is how effectively the visual and audio cues complement each other in assessing bee health. For instance, in cases where visual cues alone misclassify 36.9% of instances, audio signals correctly identify 62.6% of them. Conversely, within 29.5% misclassifications by audio cues, visual information correctly identifies 53.8%. This synergy suggests that integrating these two signals provides a more complete and precise assessment of bee health.

3) Comparison with Baseline Models: To further validate the effectiveness of the proposed CAMNN model, its performance is compared with several single-modal models using the same dataset. These baseline models include Inception v3 [54], MobileNet v2 [55], LSTM [56], and the custom CNN model that is used in bee audio detection. The CAMNN model surpasses these models in accuracy by margins of 18.0%, 18.8%, 14.8%, and 13.6% respectively. The detailed comparisons are shown in Table VIII.

V. DISCUSSION

Beekeepers often assess the health of a bee colony by analyzing both physical appearance and acoustic patterns. Healthy bees are typically energetic with a smooth, shiny outer body. However, signs like damaged wings or varroa mites may indicate problems such as wing deformities or a parasite infestation [57]. Beekeepers also listen for the colony's buzzing sounds. Healthy colonies produce a steady, uniform sound with consistent frequency and intensity, whereas stressed colonies exhibit irregular buzzing with noticeable fluctuations. Research has further decoded these acoustic signals, linking specific sounds to various colony behaviors. For instance, the distinctive piping sound made by queen or worker bees often signals critical events like swarming or queen replacement [58], [59].

The two-stage framework in the study is specifically designed to detect bees and provide only relevant data for bee health assessment. The YOLOv5 model first identifies bees with a corresponding probability and then crops them into individual images for further analysis. Notably, the model excels at distinguishing between bees and their shadows, as shown in Figure 2. Among the audio features, the Chromagram feature outperforms others in identifying bee sounds. It shows bee sounds, characterized by their specific buzzing frequencies, exhibit distinct harmonic patterns that differ significantly from non-bee sounds.

Unlike previous research that assesses beehive health using either bee images or sounds alone, this study combines both visual and audio signals. A cross-attention mechanism is employed to effectively merge the two data sequences. While self-attention processes a single embedding sequence by computing interactions within its elements, cross-attention merges two separate embedding sequences, determining how elements in visual influence those in audio. In this study,

TABLE VIII
COMPARISON WITH BASELINE MODELS

Models	Modality	Accuracy	Precision	Recall	F1-Score
CAMNN	**Visual + Audio**	**80.7%**	**82.5%**	**80.7%**	**80.8%**
Inception v3	Visual	62.7%	65.2%	62.7%	61.1%
MobileNet v2	Visual	61.9%	77.3%	61.9%	60.9%
LSTM	Audio	67.1%	71.3%	67.1%	67.6%
Custom CNN	Audio	65.9%	66.5%	65.9%	66.1%

the visual sequence functions as the query input, while the audio sequence serves as the key and value inputs to calculate attention scores. Then CAMNN selectively integrates the most relevant features from both visual and audio signals. Cross-attention has been show as important since removing it results in 4.2% accuracy drop.

The integration of visual and audio signals through the cross-attention mechanism significantly enhances bee health assessment. The CAMNN model outperforms four single-modal models. Ablation studies show that removing either visual or audio features from CAMNN leads to a marked decrease in performance. The study also demonstrates that visual and audio signals complement each other, as each signal captures cases that are misclassified by the other. Additionally, CAMNN improves model robustness by consistently achieving high F1-scores across all four bee health conditions.

Researchers in previous studies [26]–[33], [35], [38] have reported that computer vision and signal processing techniques can assess beehive health by detecting missing queens, varroa mites, and swarming by analyzing images and audio from beehive entrances. However, this study reveals that images are not sufficiently sensitive to detecting missing queens, with an accuracy of only 45.1%, and audio is less effective for identifying varroa mites, achieving just 56.8% accuracy. These findings further validate the proposed approach, demonstrating that model performance can be improved by integrating both visual and audio signals.

A comprehensive beehive monitoring system has been developed, integrating advanced technology into beekeeping. In this system, cameras and microphones are strategically positioned around beehives to capture activities. The collected data is then transmitted to an online platform via a Raspberry Pi device, as shown in Figure 8. Various models, including CAMNN, analyze the data to detect bees and assess their health. A user-friendly website enables near-real-time monitoring and evaluation of bee health. With access to real-time data and detailed insights, beekeepers can accurately identify

Fig. 8. Beehive Monitoring and Bee Health Assessment System

stressors and take timely actions to ensure colony health.

While this study provides valuable insights into applying ML models for bee monitoring, certain limitations should be noted, such as the limited sample size and restricted data sources. The study primarily focuses on Apis mellifera in California, USA, which might not fully capture the behavioral and environmental variations of other bee species or locations. Future research could aim for more extensive studies, including a wider variety of bee species, both domesticated and wild, across diverse regions. Collaborating with global bee researchers and beekeepers would lead to more comprehensive and generalizable findings.

VI. CONCLUSION

The study presents an advanced beehive monitoring and management system that enhance bee health assessments by integrating both visual and audio data. The system starts with bee object detection, identifying bees within images and audio clips. Only the data containing detected bees is then processed by the CAMNN for health assessment, which dynamically integrates visual and audio features by assigning different weights based on their significance. This approach achieves 80.7% accuracy, outperforming other four single-modal neural networks and significantly enhancing the model's robustness. The findings also highlight that audio data is more

sensitive than images in assessing hive health. With real-time monitoring and health evaluation capabilities, this system provides beekeepers with a powerful tool to promptly identify and address potential health issues within their hives.

Data Availability Statement

There are four datasets used in the study. The visual and audio datasets for object detection have been made publicly accessible on the Kaggle platform. The remaining two datasets for beehive health assessment are not currently available to the public due to considerations for intellectual property filings.

- Bee Image Detection `https://www.kaggle.com/datasets/andrewlca/bee-audio-object-detection`
- Bee Audio Detection `https://www.kaggle.com/datasets/andrewlca/bee-audio-object-detection`

References

[1] G. Bush, "How you can keep bees from becoming endangered," 2020, available: https://www.osu.edu/impact/research-and-innovation/bee-population.

[2] N. Steinhauer, "United states honey bee colony losses 2022–23: Preliminary results from the bee informed partnership," *Bee Informed*, 2023.

[3] A. Gregorc, C. Domingues, H. Tutun, and S. Sevin, "What has been done in the fight against varroa destructor: from the past to the present," *Ankara Üniversitesi Veteriner Fakültesi Dergisi*, vol. 69, no. 2, pp. 229–240, 2022.

[4] P. Hristov, R. Shumkova, N. Palova, and B. Neov, "Factors associated with honey bee colony losses: A mini-review," *Veterinary Sciences*, vol. 7, no. 4, p. 166, 2020.

[5] G. W. Otis, "Population biology of the africanized honey bee," in *The African Honey Bee*. CRC Press, 2019, pp. 213–234.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[7] C. Jones, M. Murdin, A. Rowberry, and M. May, *The Beekeeper's Guide: Building Skills and Knowledge*. Princeton University Press, 2024.

[8] P. Hristov, R. Shumkova, N. Palova, and B. Neov, "Factors associated with honey bee colony losses: A mini-review," *Veterinary Sciences*, vol. 7, no. 4, p. 166, 2020.

[9] A. Noël, Y. Le Conte, and F. Mondet, "Varroa destructor: how does it harm apis mellifera honey bees and what can be done about it?" *Emerging Topics in Life Sciences*, vol. 4, no. 1, pp. 45–57, 2020.

[10] D. Kashyap, H. Pandey, K. Jaiswal, and S. Mishra, "Fungal diseases of honey bees: Current status and future perspective," in *Recent Developments in Fungal Diseases of Laboratory Animals*, 2019, pp. 7–27.

[11] A. D. Vaudo, L. A. Dyer, and A. S. Leonard, "Pollen nutrition structures bee and plant community interactions," *Proceedings of the National Academy of Sciences*, vol. 121, no. 3, p. e2317228120, 2024.

[12] N. Danner, A. Keller, S. Härtel, and I. Steffan-Dewenter, "Honey bee foraging ecology: Season but not landscape diversity shapes the amount and diversity of collected pollen," *PLOS ONE*, vol. 12, no. 8, p. e0183716, 2017.

[13] J. H. Hunt and F.-J. Richard, "Intracolony vibroacoustic communication in social insects," *Insectes Sociaux*, vol. 60, pp. 403–417, 2013.

[14] L. Goodman, *Form and Function in the Honey Bee*. Cardiff, UK: International Bee Research Association, 2003.

[15] W. F. Towne and W. H. Kirchner, "Hearing in honey bees: detection of air-particle oscillations," *Science*, vol. 244, pp. 686–688, 1989.

[16] J. H. Hunt and F.-J. Richard, "Intracolony vibroacoustic communication in social insects," *Insect. Soc.*, vol. 60, pp. 405–417, 2013.

[17] W. H. Kirchner, C. Dreller, and W. F. Towne, "Hearing in honeybees: Operant conditioning and spontaneous reactions to airborne sound," *J. Comp. Physiol. A*, vol. 168, pp. 85–89, 1991.

[18] L. H. Field and T. Matheson, "Chordotonal organs of insects," in *Advances in Insect Physiology*, P. D. Evans, Ed. San Diego, CA: Elsevier, 1998, vol. 27, pp. 1–228.

[19] S. Bilik, T. Zemcik, L. Kratochvila, D. Ricanek, M. Richter, S. Zambanini, and K. Horak, "Machine learning and computer vision techniques in continuous beehive monitoring applications: A survey," *Computers and Electronics in Agriculture*, vol. 217, p. 108560, 2024.

[20] M. Burgess, "Acoustics australia," *Acoustics Australia*, vol. 43, no. 1, 2015.

[21] S. Pratt, S. Kühnholz, T. D. Seeley, and A. Weidenmüller, "Worker piping associated with foraging in undisturbed queenright colonies of honey bees," *Apidologie*, vol. 27, pp. 13–20, 1996.

[22] C. Collison, "A closer look: sound generation and hearing," *Bee Culture: The Magazine of American Beekeeping*, vol. 22, 2016.

[23] T. D. Seeley and J. Tautz, "Worker piping in honey bee swarms and its role in preparing for liftoff," *Journal of Comparative Physiology A*, vol. 187, pp. 667–676, 2001.

[24] M. S. Sarma, S. Fuchs, C. Werber, and J. Tautz, "Worker piping triggers hissing for coordinated colony defence in the dwarf honeybee apis florea," *Zoology*, vol. 105, pp. 215–223, 2002.

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

[26] C. Yang and J. Collins, "Deep learning for pollen sac detection and measurement on honeybee monitoring video," in *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2019, pp. 1–6.

[27] T. Sledevic, "The application of convolutional neural network for pollen bearing bee classification," in *2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 2018, pp. 1–4.

[28] M. N. Ratnayake, A. G. Dyer, and A. Dorin, "Tracking individual honeybees among wildflower clusters with computer vision-facilitated pollinator monitoring," *PLoS One*, vol. 16, no. 11, p. e0258834, 2021.

[29] M. Liu, M. Cui, B. Xu, Z. Liu, Z. Li, Z. Chu, X. Zhang, G. Liu, X. Xu, and Y. Yan, "Detection of varroa destructor infestation of honeybees based on segmentation and object detection convolutional neural networks," *AgriEngineering*, vol. 5, no. 4, pp. 1644–1662, 2023.

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

[31] A. Liang, "Effectiveness of transfer learning, convolutional neural network and standard machine learning in computer vision assisted bee health assessment," in *2022 International Communication Engineering and Cloud Computing Conference (CECCC)*. IEEE, March 2022, pp. 7–11.

[32] H. Üzen, C. Yeroğlu, and D. Hanbay, "Development of cnn architecture for honey bees disease condition," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2019, pp. 1–5.

[33] S. K. Berkaya, E. S. Gunal, and S. Gunal, "Deep learning-based classification models for beehive monitoring," *Ecological Informatics*, vol. 64, p. 101353, 2021.

[34] J. Kim, J. Oh, and T.-Y. Heo, "Acoustic scene classification and visualization of beehive sounds using machine learning algorithms and grad-cam," *Mathematical Problems in Engineering*, vol. 2021, no. 1, pp. 1–13, 2021.

[35] A. Terenzi, N. Ortolani, I. Nolasco, E. Benetos, and S. Cecchi, "Comparison of feature extraction methods for sound-based classification of honey bee activity," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 112–122, 2021.

[36] S. Cecchi, A. Terenzi, S. Orcioni, P. Riolo, S. Ruschioni, and N. Isidoro, "A preliminary study of sounds emitted by honey bees in a beehive," in *Proceedings of the Audio Engineering Society Convention 144*, Milan, Italy.

[37] "Open source beehives project," https://www.osbeehives.com/pages/about-us, accessed on 13 October 2021.

[38] A. Zgank, "Bee swarm activity acoustic classification for an iot-based farm service," *Sensors*, vol. 20, p. 21, 2020.

[39] X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li, "Audio-visual cross-attention network for robotic speaker tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2022.

[40] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 357–366.

[41] A. Khattar and S. Quadri, "Camm: cross-attention multimodal classification of disaster-related tweets," *IEEE Access*, vol. 10, pp. 92 889–92 902, 2022.

[42] D. Mrozek, R. Głorny, A. Wachowicz, and B. Małysiak-Mrozek, "Edge-based detection of varroosis in beehives with iot devices with embedded and tpu-accelerated machine learning," *Applied Sciences*, vol. 11, no. 22, p. 11078, 2021.

[43] R. Tashakkori, N. P. Hernandez, A. Ghadiri, A. P. Ratzloff, and M. B. Crawford, "A honeybee hive monitoring system: From surveillance cameras to raspberry pis," in *SoutheastCon 2017*. IEEE, 2017, pp. 1–7.

[44] "Bee health guru," https://www.beehealth.guru/, accessed on 31 October 2021.

[45] B. Z. J. L. S. Thornton, "Audio recognition using mel spectrograms and convolution neural networks," 2019, available: https://api.semanticscholar.org/CorpusID:237274283.

[46] J. S. Bridle and M. D. Brown, "An experimental automatic word recognition system," *JSRU report*, vol. 1003, no. 5, p. 33, 1974.

[47] J. B. Allen, "Short time spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, no. 3, pp. 235–238, June 1977.

[48] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[49] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, J. Poznanski *et al.*, "ultralytics/yolov5: v3.0," 2020.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] A. G. Howard, M. Zhu, B. Chen, and et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[52] C. Szegedy, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[55] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[57] R. Lucius, R. Poulin, B. Loos-Frank, R. P. Lane, C. Roberts, and R. K. Grencis, *General aspects of parasite biology*. Wiley-VCH, 2017, pp. 1–93.

[58] C. Phillips, "Telling times: More-than-human temporalities in beekeeping," *Geoforum*, vol. 108, pp. 315–324, 2020.

[59] C. Uthoff, M. Nabhan Homsi, and M. von Bergen, "Acoustic and vibration monitoring of honeybee colonies for beekeeping-relevant aspects of presence of queen bee and swarming," *Computers and Electronics in Agriculture*, vol. 205, p. 107589, 2023.

Declaration of Academic Integrity

The participating team declares that the paper submitted is comprised of original research and results obtained under the guidance of the instructor. To the team's best knowledge, the paper does not contain research results, published or not, from a person who is not a team member, except for the content listed in the references and the acknowledgment. If there is any misinformation, we are willing to take all the related responsibilities.

Names of team members: Andrew Liang

Signatures of team members:

Name of the instructor: Ms. Anuradha Datar

Signature of the instructor: 

Date: 8/22/24