Comparison of VGGish embedding and MFCC feature 1 in bee colony sound classification (#73841)

First revision

Guidance from your Editor

Please submit by **2 Dec 2022** for the benefit of the authors .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files
Download and review all files from the materials page.

1 Tracked changes manuscript(s)
1 Rebuttal letter(s)
11 Figure file(s)
5 Table file(s)
1 Other file(s)

For assistance email peer.review@peerj.com

Structure and Criteria 2

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

Clear, unambiguous, professional ~~EnglishOriginal~~English language used through ~~out.the journal~~out. The.

Intro & background to show context.Research question & relevant.& meaningful. It is stated how the Structure

fills an identified knowledge gap. discipline norm or improved for clarity.

Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

Original primary research within Scope of

well defined, relevant Literature well referenced conforms to PeerJ standards, research

Rigorous investigation performed to a Figures are relevant, high quality, wellhigh technical & ethical standard.

information to replicate.

labelled & described.

Methods described with sufficient detail &

VALIDITY OF THE FINDINGS

Impact and novelty not assessed.Conclusions are replication encouraged whereoriginal research literature is clearly supporting results. stated.

All underlying data have been provided; they are robust, statistically sound, & controlled.

well stated, linked to *Meaningful* question & limited to rationale & benefit to

Standout reviewing tips

The best reviewers use these techniques

Tip

Support criticisms with evidence from the text or from other sources

Example

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript or contact a professional editing service.

Organize by importance of the issues, and number your points

1. *Your most important issue*
2. *The next most important item*
3. *…*
4. *The least important points*

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

Comparison of VGGish embedding and MFCC feature in bee colony sound classiûcation

Nayan Di Corresp., 1, 2 , **Muhammad Zahid Sharif** 1, 2 , **Zongwen Hu** 3, 4 , **Renjie Xue** 1, 2 , **Baizhong Yu** 1, 2

[1] Anhui Institute of Optics and Fine Mechanics, Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei, China

[2] University of Science and Technology of China, Hefei, China

[3] Eastern Bee Research Institute, College of Animal Science and Technology, Yunnan Agricultural University, Kunming, China

[4] The Sericultural and Apicultural Research Institute, Yunnan Academy of Agricultural Sciences, Mengzi, China

Corresponding Author: Nayan Di
Email address: dny_yan@126.com

Background. Bee colony sound is a continuous, low-frequency buzzing sound that varies with the environment or the colony's behavior and is considered meaningful. Bees use sounds to communicate within the hive, and bee colony sounds investigation can reveal helpful information about the circumstances in the colony. Therefore, one crucial step in analyzing bee colony sounds is to extract appropriate acoustic features.

Methods. This paper uses VGGish embedding and MFCC feature generated from three bee colony sound datasets to train several machine learning algorithms to determine which acoustic feature performs best in bee colony sound recognition.

Results. The results showed that VGGish embedding performs better than or on par with MFCC feature in all three datasets.

Commented [JB1]: Define the aconyms VGGish and MFCC here

Comparison of VGGish embedding and MFCC feature in bee colony sound classification

Nayan Di[1,2,*] , Muhammad Zahid Sharif[1,2], Zongwen Hu[3,4], Renjie Xue[1,2], Baizhong Yu[1,2]

[1] Anhui Institute of Optics and Fine Mechanics, Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei,230031, China

[2] University of Science and Technology of China, Hefei,230026, China

[3] Eastern Bee Research Institute, College of Animal Science and Technology, Yunnan Agricultural University, Kunming, China

[4] The Sericultural and Apicultural Research Institute, Yunnan Academy of Agricultural Sciences, Mengzi Yunnan, China

*Corresponding Author:
Nayan Di[1]
350 Shushanhu Road, Hefei, Anhui, 230031, P. R. China
Email address: dny_yan@126.com

Abstract

Background. Bee colony sound is a continuous, low-frequency buzzing sound that varies with the environment or the ~~colony''~~ ' ~~'s~~ colony's behavior and is considered meaningful. Bees use sounds to communicate within the hive, and bee colony sounds investigation can reveal helpful information about the circumstances in the colony. One crucial step in analyzing bee colony sounds is to extract appropriate acoustic feature.

Methods. This paper uses VGGish embedding and MFCC coefficients ~~feature~~ , extracted ~~generated~~ from three bee colony sound datasets, to train several machine learning algorithms to determine which acoustic feature performs best in bee colony sound recognition.

Results. The results showed that VGGish embedding performs better than or on par with the MFCC feature in all three datasets.

Commented [JB2]: Spell out the words from which each of these acronyms is derived.

Keywords: Acoustic feature; Bee colony sound; VGGish embedding; MFCC feature; *Apis cerena* Apis mellifera? If Apis ceranae, this adds application of acoustic screening to another species of honey bee. If Apis mellifera, which is what was examined in the first draft of this study, the A. ceranae should be replaced with A. mellifera throughout the manuscript.

1. Introduction

Honey bees play an essential role in agriculture production and are almost responsible for 90% of global commercial pollination service pollination(Klein, VaissiŁre et al. 2007). As a vital node of the agriculture section, it is essential to ensure that the bee colonies can provide service.

To save human resources and reduce disturbance to bee colonies, a non-invasive method that can detect the intra-colonial condition of the hive without disturbing the colony is a consensus among researchers and practitioners (Meikle and Holst 2015). The internal environment of a honeybee hive includes sound, temperature, and humidity, which are relatively stable under certain conditions (Murphy, Magno et al. 2015). By monitoring these indicators in the hive and establishing the association between these indicators, we could learn a lot about the status of the colony (Ferrari, Silva et al. 2008, Braga, Gomes et al. 2020). Among these indicators, beehive sound is critical. Bee buzzing carries information on colony behavior and phenology. Honey bees emit specific sounds when exposed to stressors such as pest infection (Qandour, Ahmad et al. 2014), airborne toxicants~~ ~~(Zhao, Deng et al. 2021), and failing queens (Cejrowski, SzymaEski et al. 2018). Using both statistical and A.I. analysis of colony sounds, ~~Jerry~~ Bromenshenk et al., in their patents (2009) and in their review paper (2015) showed that their Artificial Intelligence (A.I.) could detect a diverse ~~developed a smartphone app(Bee Health Guru KS) which could detect~~ a diverse variety of chemicals and eight colony health variables, ~~inside beehives~~ by simply ~~put~~putting a microphone into the bottom of a beehive and recording bee colony sounds for 30 or 60 seconds. In 2019, they released a cellphone app (Bee Health Guru) that can run the diagnostic programs, record and analyze the results, and upload the data, visual inspections, and app analyses to a cloud-based site, which automatically generates a report with the GPS location shown on a map. ~~the cellphone near the beehive and recording the bee colony sound for a few seconds.~~Currently, the app is being calibrated for a variety of phone operating systems for bee sounds from around the world (www.beehealth.guru).

One of the critical phases in analyzing the bee colony sound would be extracting appropriate feature from the bee colony sound for machine learning or deep learning algorithms. Traditionally we use frequency domain or time domain feature of sound, such as soundscape indices and low-frequency signal features (Sharif, Wario et al. 2020). Mel Frequency Cepstrum

Commented [JB3]: The 2015 review is available on line and should be cited. The acoustic discrimination includes both chemical and biological endpoints.

Coefficient (MFCC) is one of the most commonly used features. It is characterized by using a set of critical coefficients to create Mel cepstrum, which makes its cepstrum more similar to the nonlinear human auditory system (Muda, Begam et al. 2010). Due to the nonlinear correspondence between Mel frequency and Hz frequency, the calculation accuracy of MFCC decreases with the increase of frequency. This characteristic makes MFCC more suitable for bee colony sound than other feature extraction methods in the past because the sound signal in the colony is concentrated in the low-frequency part (Dietlein 1985).

Commented [JB4]: This definition needs to appear the first time it is mentioned in the paper, then use MFCC. MFCC is a set of calculations that generate a coefficient. Drop the redundent and confusing phrase MFCC coefificient feature. The word feature is confusing - MFCC is a coefficient, not a feature.

Thanks to the rapid development of artificial intelligence, Convolutional Neural Net (CNN) and Recurrent Neural Networks (RNN) have been widely applied in audio recognition (Kumar and Raj 2017). Experimental results showed that the recognition method based on CNN is prior to the method based on machine learning models (Kulyukin, Mukherjee et al. 2018). Visual Geometry Group (VGG) is one of the most popular CNN models. ~~It was proposed by Simonyan~~Simonyan and Zisserman proposed it ~~and Zisserman~~ in 2014 and is named after the Visual Geometry Group (Simonyan and Zisserman 2014). VGGish is a TensorFlow definition of a VGG-like audio classification model. The VGGish model is a derivative network of the VGG network trained on a large YouTube dataset (Gemmeke, Ellis et al. 2017). Its structure is consistent with VGG11, including eight convolutional layers, five pooling layers, and three fully connected layers. Each convolutional layer uses a 3x3 convolution kernel. VGGish converts audio input feature into a semantically meaningful, high-level 128-dimensional embedding, which can be fed as input to a downstream classification model. Due to the scale and diversity of the YouTube dataset, the resulting acoustic

Commented [JB5]: The original article used YouTube derived data, this revision suggests that it wasn't used. If used, the YouTube data is likely to be for Apis mellifera, not Apis ceranae.

features are both very general and of high resolution, placing each audio sample in a high dimensional feature space that is unlikely to show ecosystem-specific bias. This 128-dimensional embedding characteristic is helpful in various identification contexts, including monitoring anomalous events in an ecosystem (Sethi, Jones et al. 2020) and sound-based disease detection (Shi, Du et al. 2019).

In this article, we contribute to the body of research on audio beehive monitoring by comparing VGGish embedding and standard MFCC statistics (?) ~~feature~~ in classifying audio samples from microphones deployed inside beehives. We tested the VGGish embedding and MFCC ~~feature~~ on three different classification tasks and compared these two-feature using four machine-~~-~~learning algorithms.

Commented [JB6]: Overall, still a very small data set. Most investigators use recordings from 100 or more colonies, splitting the data into a training and a testing group. One can not train on all samples, then test on the same set of samples.,

In particular, Section two will describe the hardware and software configuration to obtain bee colony sound and report the detail of the three bee colony datasets we used in this paper.

Section three will give the performance of VGGish embedding and MFCC ~~feature~~ coefficients in bee colony sound classification, as well as the effects of different dimensional reduction algorithms. Section four will report conclusions and a future perspective.

2. Materials and Methods

2.1 Hardware

The hardware and software systems for obtaining bee colony sound are as follows: A microphone inside the beehive (PCK200, TAKSTAR) was placed about 15cm from the bottom. The microphone has a frequency range of 30-20 kHz 30 Hz to 20 kHz (30 kHz to 20 kHz makes no sense) and a sensitivity of -35 dB. A digital sound card (UM2, BEHRINGER) was used to convert the analog signal into a digital signal. The digital signal was transmitted to a personal computer (HP 2170p, Windows 7), The software Audacity was used to record the sound, and the sound sampling rate was set to 44.1kHz, mono. Sound files were saved on the hard disk in .wav format. The hardware structure is illustrated in Figure1.

2.2 Audio data

The experiment was carried out at the Sericulture and Apiculture Research Institute of Yunnan Academy of Agricultural Sciences (23.5144N,103.4043E) from November 2020 to June 2021. The institute is located in Caoba Town, Mengzi City, Yunnan Province, China. We collected three collections of honey bees (*Apis* ~~*cerena*~~*Serena*) colony sounds and named them dataset one, two and three, respectively. A detailed description of these datasets is given below. Every bee colony lived in standard wood beehive with a queen of 10 months old. All the bee colonies were ~~are~~ healthy without any sign of attack by pests, emerging diseases, and viruses.

Commented [JB7]: Cerana or mellifera?

2.2.1 Dataset one

Dataset one contains the colony sound of three experimental groups. Each group was treated with unique odorous compounds.

Honeybees were trained with syrup solutions containing different volatile compounds to visit artificial feeding sites approximately 200 meters ~~away~~ from the hive. A feeder containing 50% sucrose solution was placed approximately five meters from the hive, and the marked foragers were caught in a glass tube at the door of the hive. The foragers were gently let out to feed on the feeder. When the foragers had eaten enough syrup, they returned to the hive after hovering over the feeder a few times. This was repeated several times, and when visited by a larger number of foragers, the feeder was slowly placed approximately 10 meters from the hive, and so on, gradually moving the feeder to the target position. When a large number of marked bees were feeding at the target distance, the sound inside the colony was recorded for 10min. Before changing the compounds added to the sucrose solution, we stopped feeding for two days, waiting for the colony to be depleted of food and odors before starting another treatment. The sound files were collected from three different colonies, each colony with two frames~~, t~~. The number of recordings and duration were shown in Table 1. In this dataset, all the colony sound files were collected during winter from November 2020 to January 2021, and very few food sources were available outside. In this way, the artificial food source we provide may be the only food ~~source~~sources for honeybees.

This dataset contains the colony sound of three experimental groups, which were treated with unique odorous compounds at a mass ratio of 0.1% in 50% (w/w) sucrose solution, sucrose solution with 50% concentration was used as blank control. The compound used were ethyl acetate and acetone. The colony sound was labeled ~~'~~" ~~blank,' 'acetone~~blank, 'acetone,~~'~~" and ~~'~~" ethyl,~~'~~" respectively~~.~~.

2.2.2 Dataset two

Dataset two collects bee colony sounds concerning the ~~queen~~~~’~~' ~~'s~~ queen's status. The object is to use the colony sound to detect whether there is a queen pupa and whether the pupa has hatched. This dataset includes honey bee sounds under three scenarios.

This work was carried out in June 2021, alternating between spring and summer. It simulated the occurrence of a new queen cell in the colony before swarming. We selected two groups of healthy and strong colonies of *Apis cerana*, each with six frames of honeybees and a normal breeding queen. In the first scenario, we caged the queen and collected colony sounds. In the second stage, we introduced a mature queen pupa into this colony. The original queen was still in the cage and, therefore, would not attack the new queen pupa. Collecting sound data began after a day. In the third stage, we opened the hive every night, checked the pupa condition, and recorded the next day after the new queen emerged. All recordings started around 11:00 am. In this way, we obtained colony sounds in three different queen states. They were labeled as~~’~~' blank,~~’~~" ~~”~~' ~~queen~~ ' queen pupa,~~’~~" or~~’~~' new queen.~~’~~'-'

2.2.3 Dataset three

This dataset contains sounds from bee colonies of different colony sizes. We investigated six bee colonies, including two colonies with two frames~~;~~, two with four, and two with six. The bee colony sound was recorded at 9:00 am for about three to ten minutes in each of the colonies, and the recorded sound files were labeled as C2~~-~~, C4~~-~~, and C6, respectively. We estimate the number of bees by weighing the colony. The weight of an empty hive is measured first, then the whole swarm of bees is shaken off into the empty hive, and the mass is measured again. The mass difference obtained is the total weight of the swarm. We estimated the number of bees per colony based on the average honeybee weight, which was 94.9mg (Table 2).

2.3 Data processing

The data processing was based on python 3.5.1 and Scikit-learn 1.0.2 (Pedregosa et al., 2012).

2.3.1 Feature extraction

VGGish Embedding. The audio sample was first split into segments of 0.96s. Each 0.96s segment was first resampled to 16 kHz using a Kaiser window, and a log-scaled Mel-frequency spectrogram was generated (96 temporal frames, 64 frequency bands). Each audio sample was then passed through CNN from Google s AudioSet project (Gemmeke, Ellis et al. 2017, Hershey, Chaudhuri et al. 2017) to generate a 128-dimensional embedding of the audio. Figure 2 shows the structure of the VGGish network and the work-flow of extracting VGGish embedding.

Commented [JB8]: Why resample and drop the frequency? Colony sounds extent into the ultrasonic range.

Mel-frequency Cepstral Coefficient (MFCC). MFCCs are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique uses two types of filters: linearly spaced and log arithmetically spaced. The signal is expressed in the Mel frequency scale to capture the phonetically important characteristics of speech. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The MFCC feature extraction procedures are as follows: windowing the sound signal, applying the FFT (Fast Fourier Transform), taking the log~~,~~ and then warping the frequencies on a Mel scale, followed by applying the inverse DCT (Discrete Cosine Transform). The 13-dimensional MFCC will be combined with the first-order difference coefficients and second-order coefficients difference to get the 39-dimensional MFCC feature.

Commented [JB9]: There is no reason to suspect that bee hearing approximates that of the human ear. Is the MFCC simply correcting the signal produced by the sound card and software like Audacity which is maximized to pick out sounds the are discernible to the human ear?

A block diagram of the structure of an MFCC processor is given in Figure 3.

2.3.2 Dimension reduction

Since the features extracted from the raw data are high-dimensional, it is not conducive to visualization. It is necessary to use the technique for dimensionality reduction to get 2D points from a high-dimensional input vector.

Commented [JB10]: Why not use an X,Y,Z, 3-D visualization?

To estimate the impact of dimension reduction, we experimented with the following dimensionality reduction algorithms: (R1) uniform manifold approximation and projection (UMAP). UMAP works by learning approximate manifolds from higher dimensional Spaces and mapping them into lower dimensional Spaces (McInnes, Healy et al. 2018); (R2) t-distributed stochastic neighbor embedding(t-SNE) (Van der Maaten and Hinton 2008). This technique is a variation of Stochastic Neighbor Embedding (Becht, McInnes et al. 2019, Diaz-Papkovich, Anderson-TrocmØ et al. 2019).

The multidimensional colony sound feature were narrowed down to two by the two algorithms. Machine learning algorithms then classify the reduced feature set.

2.3.3 Training classifiers

In this paper, we trained four well-known machine learning algorithms, namely decision tree (DT), K-nearest neighbors (KNN), support vector machine (SVM) classification, and random forests (RF). DT is a tree-structured classifier. The internal nodes represent the features. The branches represent the rules, and each leaf node represents the outcome. KNN (Altman 1992) is a supervised learning model. A majority vote classifies its neighbors in vector space, and the data is assigned to the class with the nearest neighbors. SVM classification (Hong and Cho 2008) aims to create the best decision boundary(which is called a hyperplane) that can segregate n-dimensional space into classes so that the new data point can be put into the correct category. RF(Breiman 2001) is a classifier that contains a bunch of decision trees. It takes the prediction from each tree and predicts the final output based on the majority votes of predictions from those decision trees.

We trained all four models on the same feature vectors automatically extracted from the raw audio files in three bee colony datasets. The following feature: (F1) VGGish embedding;(F2) Mel frequency cepstral coefficients (MFCC) (Davis et al., 1980) are used in training all four models. We used the mean of the test accuracy as a summary of the ~~model s~~model's performance. Then the paired Student s t-test was used to check if the difference in the mean accuracy between the two models is statistically significant.

The labeled feature were split into a training set (70%) and a testing test (30%) with the training_test_split procedure from the Python sklearn.model_selection library (Pedregosa, Varoquaux et al. 2011). All these classification models were trained with the training set on an Intel Xeon E5-2676V3@2.40 GHz x 12 processor with 64 GiB of RAM and 64-bit Windows 10.

2.3.4 Model evaluation

Classification accuracy and F1 score were used to evaluate the performance of the ML models. The classification accuracy is the percentage of correct predictions. The F1 score

Commented [JB11]: Is a student T-test the proper statistical test for data such as this?

Commented [JB12]: 79% training compared to 20 testing heavily weights the analysis model towards success with a small set of recordings.

integrates information regarding both precision and recall(Chinchor and Sundheim 1993). The balanced accuracy of the classifier on the test set was reported as an average F1 score for each class to account for sample-size imbalances among classes.

The data processing work-flow is presented in Figure 4.

3. Results

3.1 The performance of models on dataset one

3.1.1 Audio signal

Two different compounds were added separately into the sucrose solution. Figure 5 presents the log spectrogram of the bee colony sound. We can see that:1) after being treated with a compounds-sucrose solution, the low-frequency sound in the bee colony increased; 2) the bee colony sound increased more significantly when feeding with the acetone-sucrose solution than when feeding with ethyl-sucrose solution, and there was a significant increase in bee colony sound around 130hz.

Commented [JB13]: 1 drop of toluene in a multi-story, A. mellifera colony, produces and immediate, and easy to hear roaring sound. Your dosing levels are very high for an insect that can detect many odors at the parts per trillion range.

3.1.2 Dimensional reduction of audio feature

Figure 6 shows the output of VGGish embedding and MFCC ~~feature~~ dimensionality reduction in dataset one. In the two-dimensional diagram, it is evident that the MFCC feature overlaps after dimensionality reduction, while the VGG embedding can better distinguish the sound in these three situations.

3.1.3 Model evaluation

Table 3 and Table 4 summarize the results of four machine learning methods. VGGish embedding performs significantly better than the MFCC feature ($P<0.005$) and shows an advantage of about 30% over MFCC feature in all four machine learning methods, among which KNN performs best, ~~achieves~~ achieving an accuracy of 94.79%.

3.2 The performance of models on dataset two

3.2.1 Audio signal

From the log spectrogram of the bee colony sound (Figure 7), the colony with a queen pupae seemed more active than the colonies in the other two conditions. The signal around 250hz and 500hz are stronger in the sound collection Queen pupa and New queen than in the sound collection Blank.

3.2.2 Dimensional reduction of audio feature

Compared with the MFCC dimensionality reduction diagram (Figure 8), the scatter plot of VGG embedding after dimensionality reduction has less overlap.

3.2.3 Model evaluation

The MFCC feature performs slightly better than VGGish embedding and shows an advantage of about 4 percent in all four machine learning methods(Table 3, Table 4)~~, but~~. Still, the difference was not statistically significant (P>0.05). Moreover, KNN performed best, and achieved an accuracy of 90%.

3.3 The performance of models on dataset three (Identifying colony size)

3.3.1 Audio signal in dataset three

This dataset includes bee colony sounds from 3 different colony sizes: C2)bee colony size of about 7500 work bees; C4) bee colony size of about 11000 work bees; C6) bee colony size of of about 17000 work bees. Figure 9 presents the log spectrogram of the bee colony sound signals of in this dataset.

3.3.2 Dimensional reduction of audio feature

The output of UMAP (Figure 10) exhibits the VGGish embedding and MFCC ~~feature~~ of colony sound in dataset three.

3.3.3 Model evaluation

The accuracy of four machine learning models using different colony sound features on dataset three is shown in Table 3. VGGish embedding has an advantage over the MFCC feature of about 20 percent in all four machine learning methods, and the difference was statistically significant (P<0.05). Moreover, KNN performed best and achieved an accuracy of 91%.

3.4 The influence of different dimensionality reduction methods

~~In order t~~To test the effects of different dimensionality reduction algorithms on the accuracy of the models, w~~. W~~e have chosen two dimensionality reduction algorithms, namely UMAP and t-SNE.

Figure 11 exhibits the results of dimensionality reduction of dataset one using the t-SNE algorithm~~, ~~compared with the output of the UMAP algorithm(Figure 6), UMAP performs better than t-SNE feature in separating bee colony sounds. Table 4 shows the accuracy of four machine learning methods trained by ~~two dimension~~two-dimension factors obtained by UMAP and t-SNE. The original sound feature used by those dimensional reduction algorithms were the MFCC feature. The results show that UMAP performs better than t-SNE in almost all datasets and all machine learning methods.

4. Discussion

Hive monitoring based on colony sound has made a lot of research achievements in recent years (Terenzi, Cecchi et al. 2020) and has become increasingly popular with~~in~~ many international companies such as Arnia, Bee Hero, Nectar, and Broodminder (https://www.umt.edu/bee/monitoringconference_2020/).

In this paper, We compared the performance of VGGish embedding and MFCC feature of bee colony sound in four classification algorithms. The result in Table 3 indicated that all four classification algorithms could generate prediction accuracy percentages that are better than ~~'' chance''~~ 'chance' based percentages. In all classification methods, the VGGish feature can guarantee more than 80% testing accuracy, among which KNN has the best performance of 94%. The testing accuracy of the MFCC feature varies a lot between different datasets. In datasets one and three, the MFCC could only achieve an accuracy of about 69%, while in dataset two, it achieved an accuracy of 90%. Results (Table 3, Table 5) show that the difference between the two features in datasets one and three is statistically significant($P<0.005$). At the same time, in dataset two, there is no~~t any~~ significant difference between the two models($P>0.005$).

We confirm that the VGGish embedding applies to bee colony sound classification and performs more stability than the MFCC feature among different datasets. This may be due to the MFCC ~~being~~ is highly dependent on data and feature, which causes weak generalization ability due to insufficient bee colony data and the similarity of bee colony sound. The VGGish network is trained on a more extensive and general Audio set, which means a better generalization ability.

Our results suggest that different compounds do lead to different responses in the bee colony (Figure 6, Table 3, Table 4), ~~it~~ which further confirms the results of previous studies(Bromenshenk, Henderson et al. 2009, Sharif, Wario et al. 2020, Zhao, Deng et al. 2021, Yu, Huang et al. 2022), and ~~moreover,~~ verifies the applicability of VGGish embedding for the classification of bee colony sounds. As seen from the log spectrum of bee colony sounds (Figure 5), the acetone-sucrose solution and acetone ethyl-sucrose solution would agitate the colony compared to the sucrose solution. The ~~low-~~low-frequency amplitude was much larger when treated with acetone than when treated with sucrose solution. This may be due to the fact that acetone stimulates bee colonies more strongly than ethyl acetate at the same concentration, and low concentrations of ethyl acetate were mildly attractive to bees (Schmidt and Hanna 2006).

The MFCC feature performs better in dataset two(Table 3, Table 4). This may be because of the fact that the sound changes fundamentally during bee swarming (Michelsen, Kirchner et al. 1986). Thus, it is easier for the standard MFCC ~~feature~~ to capture the character in colony sounds. Dataset three is ~~relatively small~~small. The total duration of sound in dataset three is less than one hour, and the machine learning models trained by the VGGish embedding could still achieve an accuracy of around 90%, which may be because the VGGish could better capture the distinctions among the ~~dataset~~datasets. We have compared two different dimensionality reduction algorithms(Figure 11, Table 5), and UMAP performs better than the t-SNE in ~~almost every~~every situation. The secret of UMAP lies in its ability to infer local and global structures while maintaining relative global distances in low-dimensional space. The result also shows that UMAP performed better in separating different colony sounds.

In summary, the results of this paper indicate that the combination of VGGish embedding and the KNN method has achieved the highest accuracy on the testing set of all three datasets (Table 3, Table 4, Table 5).

Several ways in which this research can be improved are given below:

1) Beehive sound samples are ~~relatively few~~few, and only one type of microphone is used for collecting the sound, which causes a lack of data diversity and affects the ~~model' 's~~ model's generalizability. A more comprehensive data set must be attained in future work to train the system and improve the ~~model' 's~~ model's generalizability.

2) Expand the application of the model: in this study, we applied VGGish embedding in the ~~classification~~classifications of three datasets. Beehive sound can be influenced by many other factors, such as the invasion of natural enemies and parasites. Subsequent studies can check how VGGish embedding performs in these scenarios.

Acknowledgments

The authors would like to thank the Sericultural & Apicultural Research Institute for permission to conduct this study and ~~for~~ help during data collection. We also thank Xuewen Zhang, Chuntao Zhou, Chunhui Miao, and Xinqiu Huang, who have generously shared their time and expertise.

References

Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician **46**(3): 175-185.

Becht, E., et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP." Nature biotechnology **37**(1): 38-44.

Braga, A. R., et al. (2020). "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies." Computers and Electronics in Agriculture **169**: 105161.

Breiman, L. (2001). "Random forests." Machine learning **45**(1): 5-32.

Bromenshenk, J. J., et al. (2009). Honey bee acoustic recording and analysis system for monitoring hive health, Google Patents.

Cejrowski, T., et al. (2018). Detection of the bee queen presence using sound analysis. Asian Conference on Intelligent Information and Database Systems, Springer.

Chinchor, N. and B. M. Sundheim (1993). MUC-5 evaluation metrics. Fifth Message Understanding Conference (MUC 5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993.

Diaz-Papkovich, A., et al. (2019). "UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts." PLoS genetics **15**(11): e1008432.

Dietlein, D. G. (1985). "A method for remote monitoring of activity of honeybee colonies by sound analysis." Journal of Apicultural Research **24**(3): 176-183.

Ferrari, S., et al. (2008). "Monitoring of swarming sounds in bee hives for early detection of the swarming period." Computers and electronics in agriculture **64**(1): 72-77.

Gemmeke, J. F., et al. (2017). Audio set: An ontology and human-labeled dataset for audio events. 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE.

Hershey, S., et al. (2017). CNN architectures for large-scale audio classification. 2017 ieee international conference on acoustics, speech and signal processing (icassp), IEEE.

Hong, J.-H. and S.-B. Cho (2008). "A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification." Neurocomputing **71**(16-18): 3275-3281.

Klein, A.-M., et al. (2007). "Importance of pollinators in changing landscapes for world crops." Proceedings of the royal society B: biological sciences **274**(1608): 303-313.

Kulyukin, V., et al. (2018). "Toward audio beehive monitoring: Deep learning vs. standard machine learning in classifying beehive audio samples." Applied Sciences **8**(9): 1573.

Kumar, A. and B. Raj (2017). "Deep cnn framework for audio event recognition using weakly labeled web data." arXiv preprint arXiv:1707.02530.

McInnes, L., et al. (2018). "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426.

Meikle, W. and N. Holst (2015). "Application of continuous monitoring of honeybee colonies." Apidologie **46**(1): 10-22.

Michelsen, A., et al. (1986). "The tooting and quacking vibration signals of honeybee queens: a quantitative analysis." Journal of Comparative Physiology A **158**(5): 605-611.

Muda, L., et al. (2010). "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083.

Murphy, F. E., et al. (2015). b+ WSN: Smart beehive for agriculture, environmental, and honey bee health monitoring Preliminary results and analysis. 2015 IEEE Sensors Applications Symposium (SAS), IEEE.

Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.

Qandour, A., et al. (2014). "Remote beehive monitoring using acoustic signals."

Schmidt, J. O. and A. Hanna (2006). "Chemical nature of phagostimulants in pollen attractive to honeybees." Journal of Insect Behavior **19**(4): 521-532.

Sethi, S. S., et al. (2020). "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set." Proceedings of the National Academy of Sciences **117**(29): 17049-17055.

Sharif, M. Z., et al. (2020). "Soundscape indices: new features for classifying beehive audio samples." Sociobiology **67**(4): 566-571.

Shi, L., et al. (2019). "Lung sound recognition algorithm based on vggish-bigru." IEEE Access **7**: 139438-139449.

Simonyan, K. and A. Zisserman (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

Terenzi, A., et al. (2020). "On the importance of the sound emitted by honey bee hives." Veterinary Sciences **7**(4): 168.

Van der Maaten, L. and G. Hinton (2008). "Visualizing data using t-SNE." Journal of machine learning research **9**(11).

Yu, B., et al. (2022). "A matter of the beehive sound: Can honey bees alert the pollution out of their hives?" Environmental Science and Pollution Research: 1-11.

Zhao, Y., et al. (2021). "Based investigate of beehive sound to detect air pollutants by machine learning." Ecological Informatics **61**: 101246.

Figure 1

The hardware system used to obtain bee colony sound.

The microphone is placed inside the beehive~~, then t~~. The sound signal captured by the microphone is converted to a digital signal by the digital sound card, then transmitted to the PC and saved on a hard disk for further analysis.

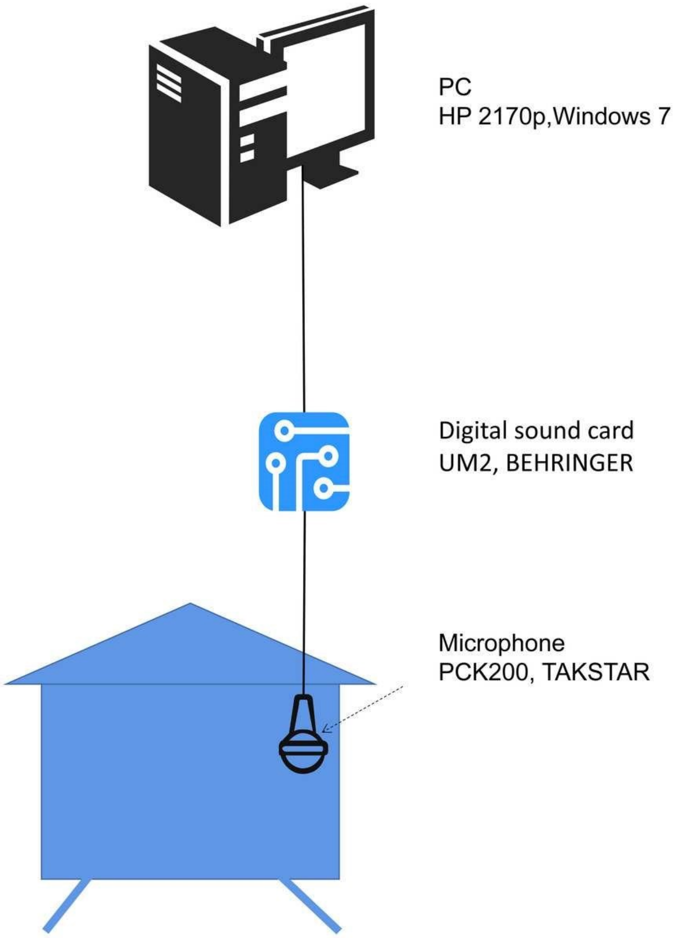

PC
HP 2170p,Windows 7
Digital sound card
UM2, BEHRINGER
Microphone
PCK200, TAKSTAR

Figure 2

An overview of the structure of the VGGish network.

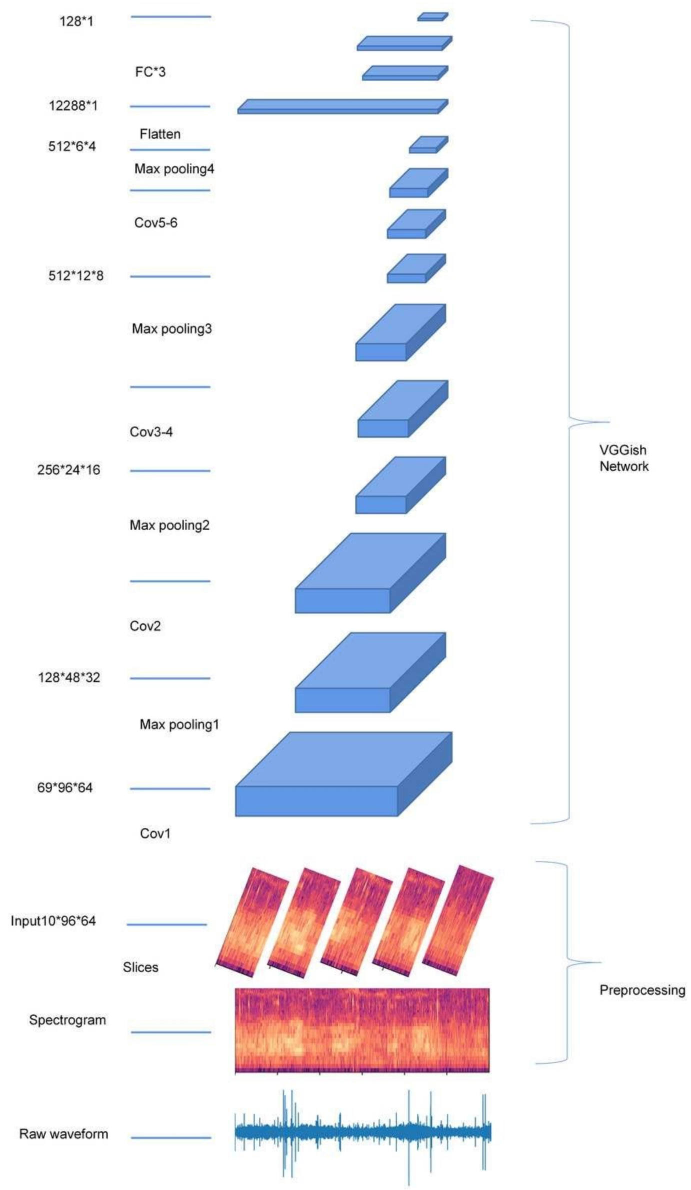

128*1
FC*3
12288*1
Flatten
512*6*4
Max pooling4
Cov5-6
512*12*8
Max pooling3
Cov3-4
256*24*16
Max pooling2
Cov2
128*48*32
Max pooling1
69*96*64
Cov1
VGGish Network
Input10*96*64
Slices
Spectrogram
Preprocessing
Raw waveform

Figure 3

Block diagram of the MFCC processor.

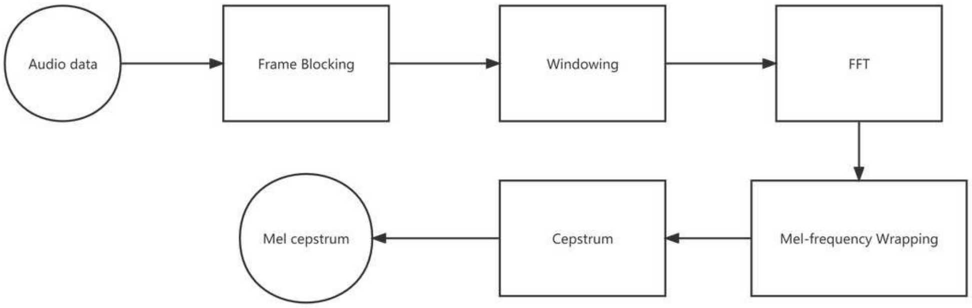


Figure 4

Overview of the approach adopted for the acoustic classiûcation of beehive sounds workûow (work-flow).

The original audio ûles (files) (.wav format) containing recordings of beehive sounds were manually classiûed (classed) into corresponding scenarios. Then, the MFCC and VGGish embedding were used to extract the audio features, respectively. Dimensionality reduction was performed using the UMAP method for the two sets of feature data. After that, the resulting data set was split into 70% for the training/development set and 30% for the testing data set. ~~The~~ Finally, the test data set was used to evaluate the performance of the classiûers (classifiers) in correctly assigning the beehive sound to the respective scenario.

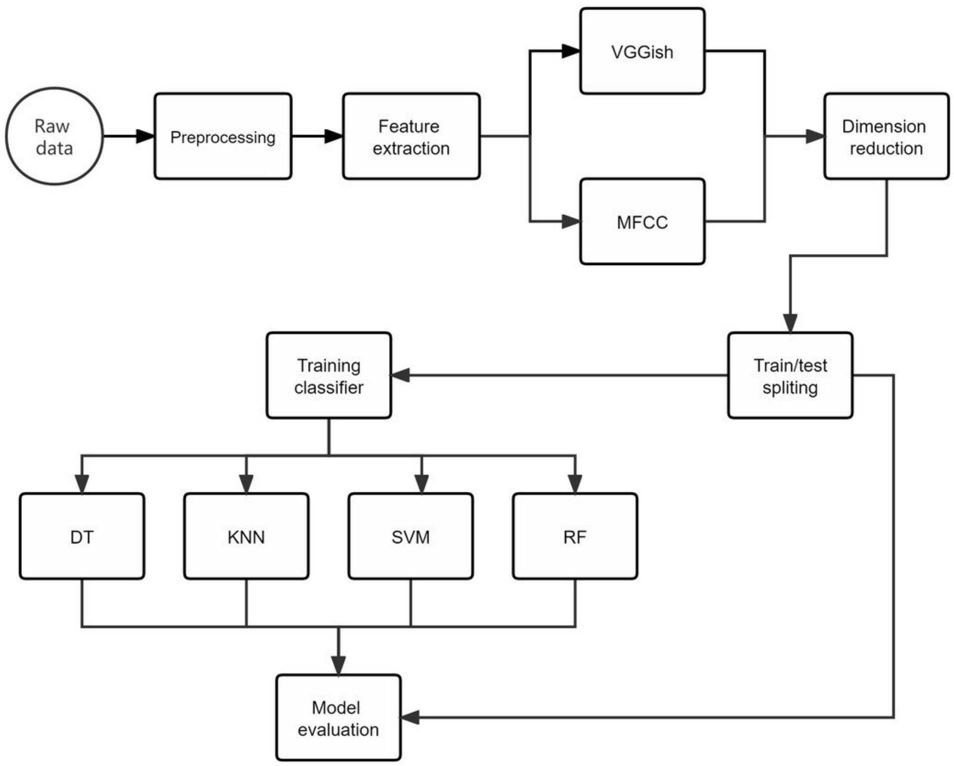

Raw data
Preprocessing
Feature extraction
VGGish
MFCC
Dimension reduction
Train/test spliting
Training classifier
DT
KNN
SVM
RF
Model evaluation

Figure 5

Log spectrum of bee colony sounds from dataset one.

Left: Acetone(treated with acetone-sucrose solution); Middle: Ethyl(treated with ethyl acetate-sucrose solution); Left: Blank(treated with sucrose solution).

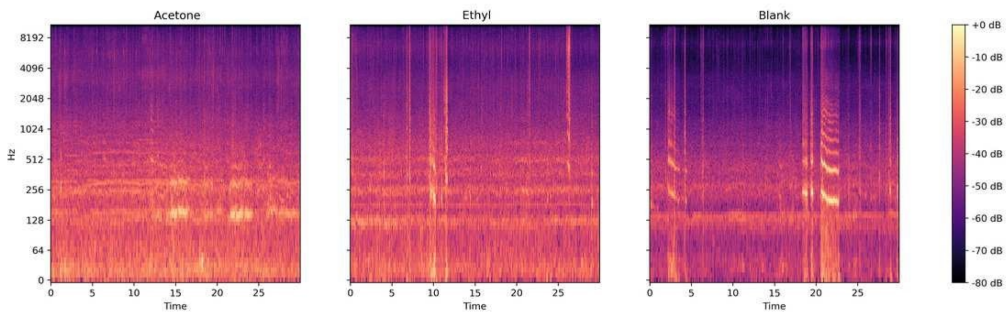
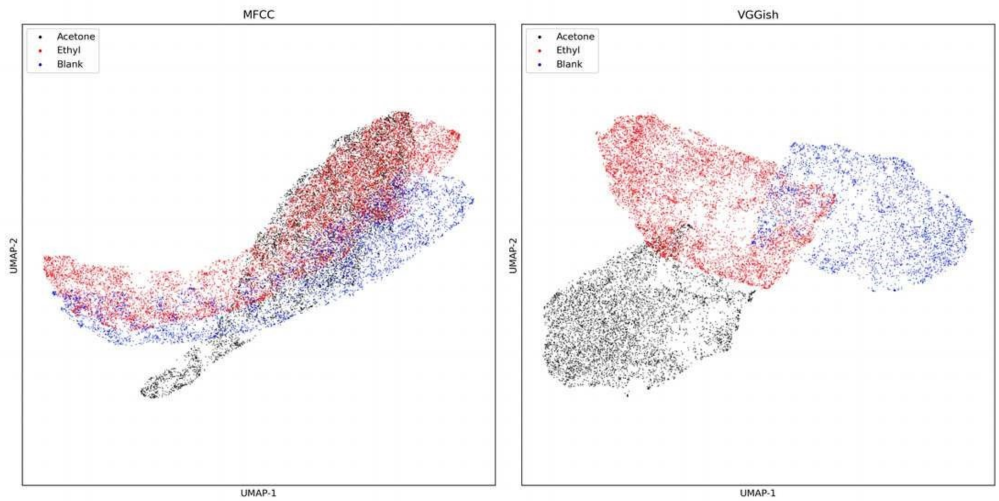


Figure 6

UMAP dimension reduction of sound features from dataset one.



Log spectrum of bee colony sounds of dataset two.

Left: Normal situation; Middle: Queen pupa inside colony; Left: New queen emerged(two queens in the colony).

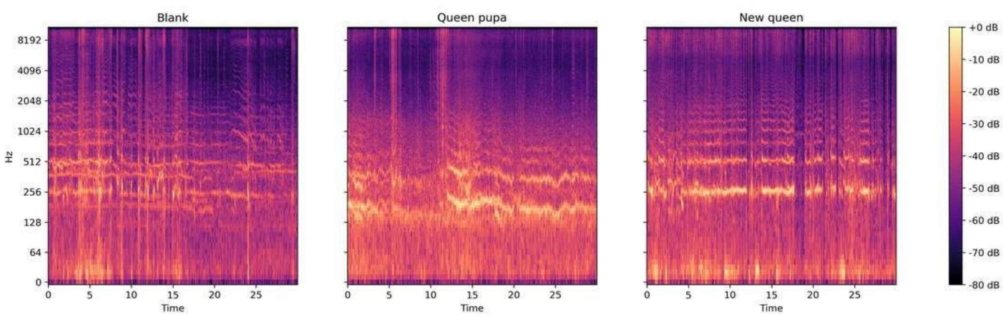
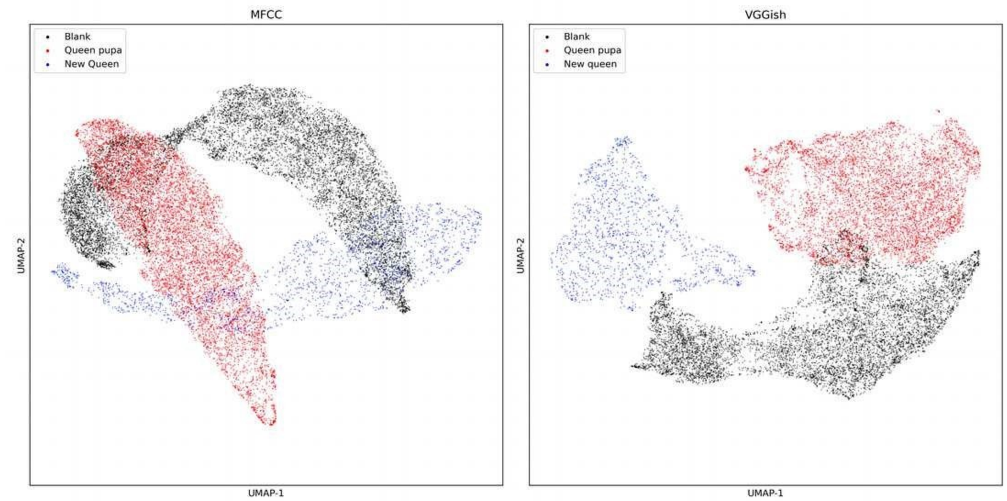


Figure 7

UMAP dimension reduction of sound features from dataset two.



Log spectrum of bee colony sounds for dataset three.

Left: Colony size of around 7500 bees(C2); Middle: Colony size of around 11000 bees(C4); Right: Colony size of around 17000 bees(C6).

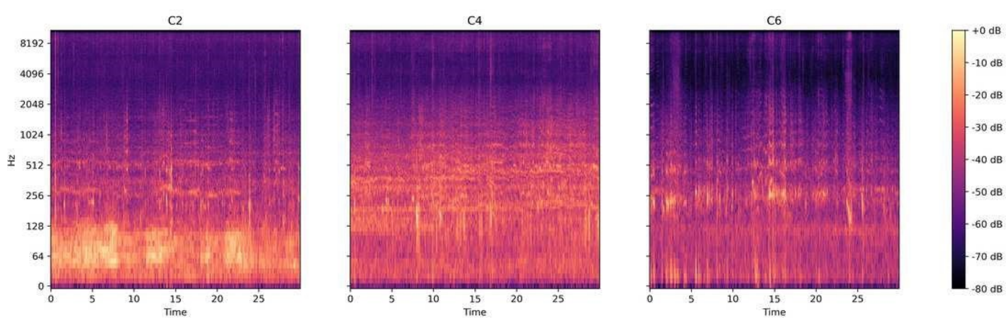
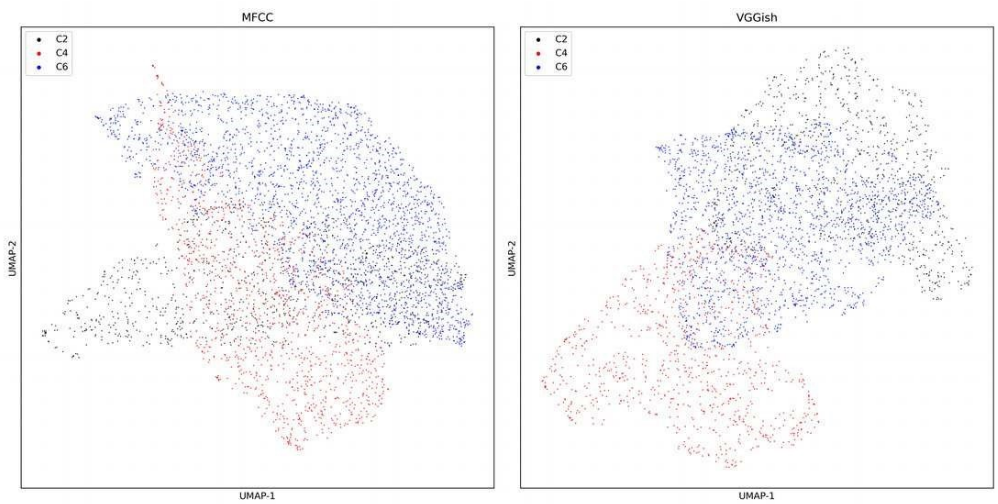


Figure 8

UMAP dimension reduction of sound features for dataset three.



MFCC features of three datasets after t-SNE dimensionality reduction.

Left: MFCC feature using t-SNE dimensionality reduction on dataset two; Middle: MFCC feature using t-SNE dimensionality reduction on dataset one; Right: MFCC feature using t-SNE dimensionality reduction on dataset three.

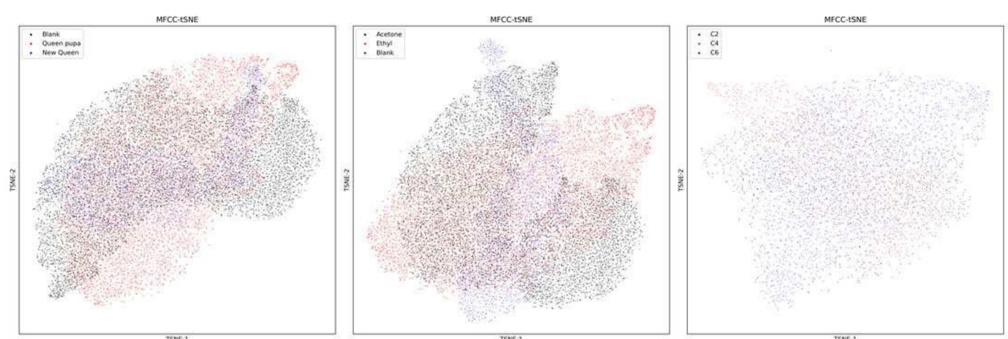


Table 1

The size of each colony used in dataset three.

"N frames" denotes the number of frames in the colony; "Total bee weight(Kg)" represents the total weight of each colony; <N worker bees= denotes the approximate number of worker bees in each colony.

(on next page)

Colony	N frames	Total bee weight(Kg)	N worker bees
1#	2	0.723	7619
2#	2	0.685	7218
3#	4	1.010	10643
4#	4	1.095	11538
5#	6	1.650	17387
6#	6	1.580	16649

Table 2

An overview of the datasets collected in order to identify compounds in nectar and queen's presence.

<Scenario= "N recordings" denotes the number of individuals with buzzing sounds recorded; "Total duration" represents the total recording time in each case; N colonies denotes the number of colonies in which we recorded sounds; N frames represent the colony size.

(on next page)

Datasets	Scenario	N colonies	N frames	N Recordings	Total Duration
Dataset one Identify compounds		3	2	6	50min
	Acetone	3	2	9	90min
	Ethyl acetate	3	2	11	111min
Dataset two Identify queen state	Blank	2	6	12	131min
	New queen pupa	2	6	9	101min
	New queen	2	6	3	23min
Dataset three Identify colony size	C2	2	2	2	12min
	C4	2	4	2	15min
	C6	2	6	2	29min

1

Table 3

Accuracy of machine learning models using diûerent (different) colony sound features on three colony sound datasets

(on next page)

Datasets	Dataset 1				Dataset 2				Dataset 3			
Algorithm	KNN	DT	RF	SVM	KNN	DT	RF	SVM	KNN	DT	RF	SVM
VGGish	94.79%	93.45%	94.43%	91.56%	86.58%	85.14%	85.94%	81.46%	91.08%	88.81%	89.23%	89.15%
MFCC	69.09%	66.28%	69.17%	68.29%	90.48%	88.45%	89.95%	87.25%	66.04%	65.78%	65.13%	68.05%

1
Table 4

F1-score of machine learning models using diûerent (different)colony sound features on three colony sound datasets

(on next page)

Datasets	Dataset 1				Dataset 2				Dataset 3			
Algorithm	KNN	DT	RF	SVM	KNN	DT	RF	SVM	KNN	DT	RF	SVM
VGGish	94.79%	93.45%	94.42%	91.55%	86.58%	85.17%	85.93%	81.49%	91.06%	88.82%	89.21%	89.03%
MFCC	68.24%	66.32%	68.49%	65.26%	90.13%	88.44%	89.63%	85.41%	65.73%	65.74%	64.80%	65.09%

1
Table 5

Comparison of diûerent (different) dimensionality reduction methods

(on next page)

Datasets	Dataset 1				Dataset 2				Dataset 3			
Algorithm	KNN	DT	RF	SVM	KNN	DT	RF	SVM	KNN	DT	RF	SVM
	69.09%	66.28%	69.17%	68.29%	90.48%	88.45%	89.95%	87.25%	66.04%	65.78%	65.13%	68.05%
t	51.62%	54.85%	55.07%	56.63%	62.64%	65.38%	66.83%	66.42%	52.24%	55.04%	57.45%	60.18%