Beehive Sound: Can It Indicate The Response Of Honey Bees To Chemicals In Or Out Of The Beehive?

Baizhong Yu
Hefei Institutes of Physical Science Chinese Academy of Sciences: Chinese Academy of Sciences Hefei Institutes of Physical Science
Xinqiu Huang
Yunnan Academy of Agricultural Sciences
Muhammad Zahid Sharif
Hefei Institutes of Physical Science Chinese Academy of Sciences: Chinese Academy of Sciences Hefei Institutes of Physical Science
Xueli Jiang
Hefei Institutes of Physical Science Chinese Academy of Sciences: Chinese Academy of Sciences Hefei Institutes of Physical Science
Nayan Di
Hefei Institutes of Physical Science Chinese Academy of Sciences: Chinese Academy of Sciences Hefei Institutes of Physical Science
Fanglin Liu (flliu@ipp.ac.cn)
Hefei Institutes of Physical Science Chinese Academy of Sciences: Chinese Academy of Sciences Hefei Institutes of Physical Science

Research Article

Keywords: Honey bees, Air contamination, Beehive sound, Acetone, Ethyl acetate, Machine Learning

Posted Date: March 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1302604/v1

Abstract

Honey bees (*Apis* spp.) are widely used as biological indicators of environmental changes. Recently, bees have been explored by researchers to monitor air contamination by listening to their hive sound. However, no study has determined whether beehive sound reflects the responses of bees to in-hive or out-hive chemicals. In this study, we conducted a feeding experiment to address this. First, we fed colonies with pure syrup (PS), syrup containing acetone (SA) or syrup containing ethyl acetate (SE) to collect beehive sound to establish multiple classifications using machine learning (ML) models. Then, we orderly fed colonies with PS, followed by SE, SA and PS. Next, we fed colonies in PS, SA, SE and PS order. Eventually, we evaluated the recall and precision of the model in detecting each syrup type. The result built on orderly feeding had a recall of 99%, 80%, 30%, 53% in detecting PS, SE, SA, PS, respectively. In the reverse feeding experiment, the ML model has a recall of 99%, 89%, 37% and 44% in detecting PS, SA, SE, and PS, respectively. Because the collected syrup in the two orderly feeding sessions was not removed from the frames during the experiment, the results indicate that beehive sound responds to chemicals in or out of the beehive.

Introduction

Insect pollinators, including honeybees, play an important role in ecosystem health maintenance and global food security. In fact, approximately 87.5% of flowering plants rely on bees for reproduction (Ollerton et al., 2011). However, climate change and human activities, especially the use of pesticides and herbicides (Dicks et al., 2020), have been shown to cause massive bees' decline (Potts et al., 2010). It is very crucial to monitor environmental change and control the usage of pesticides and fertilizers.

Biomonitoring uses the reactions of individuals, populations or communities to environmental pollution or changes to clarify the environmental pollution situation and provides a biological basis for monitoring and evaluating environmental quality (Wang et al., 2010). In the existing research, many representative plants and animals have been used for biological monitoring. For example, flathead mullet, *Mugil cephalus*, is a widely used biomonitor worldwide within coastal zones (Waltham et al., 2013), earthworms are considered to be an appropriate biological monitoring animal for determining the ecological hazards of contaminated soil (Xu et al., 2009), and mosses have been shown to be an ideal and reliable biomonitor and used as an indicator for the detection of trace metal pollution in the atmosphere (Mahapatra et al., 2019). Compared with traditional physical and chemical monitoring methods, biological monitoring has many advantages.

Honeybees are one of the most widely used insects in agricultural production. Humans manage honeybees for their production of honey and pollination services (Themudo et al., 2020). In recent studies, honeybees were used as biomonitors to detect the distribution of contaminants. Usually, honeybees explore an area within 2 kilometers of the hive, taking nectar, pollen and water. In the course of the foraging process, ecological samples from the environment are collected into the hive. Systematic variations in compounds and trace element concentrations of honey from gathered from various colonies

closely reflect anthropogenic land-use activities (Smith et al., 2019). Researchers monitored pesticides, heavy metals and volatile organic compounds in the environment by detecting the composition of substances in bees and honey (Balayiannis et al., 2008; Cochard et al., 2021; Sari et al., 2021). In addition, honeybees have extraordinary olfactory capabilities and are highly sensitive to toxic substances. The honey bee genome sequence revealed that 170 odorant receptors were annotated in the bee (Robertson & Wanner, 2006). When exposed to low-dose insecticides, honeybees may be induced to produce nitric oxide as a defense mechanism (Bartling et al., 2021). Luo et al. (2021) also showed that glyphosate at a 1/2 recommended concentration significantly affected the olfactory learning and memory ability of honeybees.

The beehive sound contains some information about the state of the bee colony. Honeybees produced unique acoustic signals upon exposure to various environmental stress. For example, queenless colonies produce sounds different from those with a queen. Although the human ear can detect changes in beehive sound without assistance, people cannot identify the sound and what caused the problem (Bromenshenk et al., 2009). Pérez et al. (2016) analysed sound patterns to evaluate changes in beehive behaviour. In this work, they propose monitoring beehive health although the evolution of the frequency spectrum sampled in four temporal windows. Based on the development of precision beekeeping and the establishment of an IoT-based beehive monitoring system (Zacepins et al., 2015; Tashakkori et al., 2021), the way to collect sound information about bee colonies from beehives has matured. In recent studies, the sound of hives has been used as an indicator to assess the status of hives. Through both support vector machines and convolutional neural networks, Nolasco & Benetos (2018) explored some aspects of hive sound recognition systems with ML methods. Lasso logistic regression and singular value decomposition are currently available to analyse the sound of hives to identify sound patterns in queenless bee colonies (Robles-Guerrero et al., 2019). Kawakita et al. (2019) recorded hive sounds of *Apis cerana japonica* over 24 h, and they found that *Apis cerana japonica* hissing had unique temporal patterns. Cejrowski et al. (2020) also presented a method that uses Mel Frequency Cepstral Coefficients (MFCCs) features and an SVM classifier to define the start and the end of the presumed bees-night in summer. In addition to MFCCs, soundscape indices are considered to be a valid set of features for acoustic analysis of a beehive, and call the attention of research community to further employ them in bioacoustics-related investigations (Sharif et al., 2020).

Regarding the stress of the volatile compound, Zhao et al. (2021) proved that the chemical information in the air surrounding the beehives can be provided qualitatively by beehive sound analysis. In the Zhao experiment, honeybees were exposed to chemical compounds in beehives. They assumed that the air conditions in the beehive were consistent with the external conditions. In this way, honeybees were thought to recognize the types of compounds in the air. However, there are two possibilities for this assumption. One is that honeybee colonies were affected by compounds in the beehive. Another explanation is that honeybees detect chemicals and caution the colony about chemical information.

In this study, acetone and ethyl acetate were used to influence beehive sound. Each of them is a volatile chemical with irritating odors. The major objective of the research is collecting beehive sound, which is

influenced by compounds, and then filtering important features of sound by data mining for training ML models to identify the syrup category. In this way, we determine whether the beehive sound about incoming new chemicals changed due to honeybees exposed to other chemicals, which remained in beehives.

Methods And Materials

Experiments design

In this study, *Apis cerana* was used as the experimental object. The research was carried out in a mulberry field at the Sericultural and Apicultural Research Institute Yunnan Academy of Agricultural Sciences in Honghe Hani and Yi Autonomous Prefecture, Yunnan Province, China (Fig. 1). Four bee colony hives, each hive with approximately 9000 bees, were placed in an open area in a mulberry field (103.39°E, 23.52°N). Acetone and ethyl acetate were purchased from Sinopharm Chemical Reagent Co., Ltd. (Shanghai, China). Sound acquisition, transmission, and storage devices use the same iPhone-8 model with 4G mobile network services.

Data Acquisition

The iPhone-8 was positioned above the beam cabinet and separated from bees by steel net, which was used to prevent interference from bees. Audios of beehive sound were recorded by the recording function of iPhone-8 in the default settings. The sound data are recorded in mono and MPEG-4 file formats. The audio sample rate is 22 kHz with 16 bit resolution.

To collect accurate audio data from the hive, we trained honeybees to the feeding station, which is 30 meters away from the beehive (Fig. 1), before the experiment. During the experiment, a sugar feeder was placed in the feeding station, which contained 500 grams of sugar water, in which the weight ratio of sugar was 50%. The treatment of syrup with a chemical compound added acetone or ethyl acetate to the weight ratio of 0.1%. In this way, syrup was divided into three categories: PS, SA and SE. Before the experiment, we randomly fed colonies PS, SA or SE to collect beehive sound for data mining and building ML models. The fragments of each sound last approximately 30 minutes. Then, we collect another set of sound data in the following experiments and use it to test the ML model predictions.

In the first step of the experiment, PS was placed in the feeder at 8 in the morning. Figure 2(a) presents the process of recording sound of each type of syrup. In each process of collection, we did not start an audio recording until the bees had foraged to the feeder for 10 minutes. The process of collecting audio data was conducted for at least 30 minutes without interruption in each hive. Then, we removed the feeder in the next 20 minutes to restrict the bees foraging or visits the feeding station. In the second step, we placed SE in the feeder and replicated the aforementioned process. In the next step, we placed SA in feeder. At last, PS was used to collect sound.

When the first phase was completed, we stopped the experiment for at least three days to dissipate the chemical compounds from beehives. Then, we started the next phase at the same time on another day. In the second phase, we changed the order between SA and SE. Next, we replicated the process of first phase. The specific process of the experiments is shown in Fig. 2(b).

Data Mining

In modern society, big data are generated and stored every day, which promotes the development of data mining approaches. Data mining usually refers to the process of finding hidden information from a mass of data (Li et al., 2017). With the development of computer technology, researchers apply intelligent methods to extract data through data mining (Barati et al., 2011). At present, data mining has been widely used in many fields, such as medicine, ecology and genomics (Chen et al., 2011; Vizcaino et al., 2014; Han et al., 2020). In addition, classification algorithms are used for data mining (Yasodha & Prakash, 2012; Anitha & Kaarthick, 2021). In this study, the steps of data mining were divided into the following stages: (1) preprocessing of data; (2) filtering and sorting of importance of features; and (3) validation of the classification model.

After the audio was captured, all of the audio files were moved to a computer, and each of the original files was converted from MPEG-4 to waveform (wav) by Python. After that, audio of all lengths was cut to a 30-min audio file. Then, they were divided into 10-s samples without overlap. Next, the R programming language (TeamRCore, 2013) was used to extract common signal characteristics, which included the low-level signal features, 13 MFCCs (Nolasco & Benetos, 2018) and 12 chroma vectors (CVs) (Müller et al.,2005), from all 10-s samples.

In these features, MFCCs are one of the most widely used feature extraction means of sound. Through the parameter set, which is based on a mel-frequency cepstrum, Davis & Mermelstein (1980) demonstrated the superior performance of MFCCs in the recognition of short-term audio spectra. In the field of voiceprints, MFCCs stand out in terms of artificial features. From speech recognition to bridge health monitoring (Lin et al., 2014; Mei et al.,2019), MFCCs have been widely used in scientific research. As described by Logan (2000), after inputting the sound signal, the following steps were taken: pre-emphasizing, framing, windowing, carrying fast Fourier transform (FFT), mel-frequency warping, calculating the filter bank and finally taking the discrete cosine transformation (DCT).

Before building the ML model, random forest (RF) was used to estimate the feature performance and the importance of different features in modeling. In the rest of the data, 80% of the 10-s samples were randomly chosen to be the training group, and the other samples were used as the test group. The features of audio samples were grouped into PS, SA and SE. We built a binary classification model based on characteristics of the audio sample in the training group by RF. Through the trained RF model, we evaluated the importance of every predictor variable by the mean decrease in accuracy (MDA) and mean decrease in Gini (MDG) (Calle & Urrea, 2011). For subsequent calculations, we chose MDA as the

indicator to predict the importance of features. After unimportant features were discarded, the remaining features were used for the establishment and prediction of subsequent training ML models.

Building Models

The k-nearest neighbor (KNN) is a basic classification and regression method (Cover & Hart, 1967; Tan et al.,2006). In December 2006, KNN was identified as one of the top 10 data mining algorithms by the IEEE International Conference on Data Mining (ICDM) (Wu et al.,2008). The input of KNN is the test data and training sample dataset, and the output is the category of the test sample. During the test, the distance between the test sample and all training samples is calculated and forecasted by the majority vote based on the category of the nearest K training sample. The three elements of the KNN are distance measurement, k size, and classification rules. Recently, some improved algorithms based on the KNN have been used in a variety of studies, such as clustering for large-scale data and human activity recognition (Chen et al., 2019; Tan et al.,2021). In this paper, KNN was used as a supervised learning model to classify data.

RF is a combination of tree predictors. In the RF, each tree relies on the value of a random vector sampled independently, and all trees in the forest are distributed in the same way (Breiman, 2001). In fact, each decision tree is a classifier. Therefore, for an input sample, the classification results are as many as the trees in the RF. For this reason, RF integrates all classified results, specifying the category with the most votes as the final output. Because of this feature, RF has been widely used in various research fields, such as computational toxicology (Mistry et al., 2016) and visual image classification (Xu Y. et al., 2018). Based on RF, Xia et al. (2018) proposed a method of detecting acoustic events using contextual information and bottleneck characteristics. In this study, the trained RF model categorized the data of each audio sample in the test group to determine which category it is the most likely to belong to.

SVM is a classification technique based on the optimal margin in ML. It was proposed as a training algorithm to maximize the margin between the training mode and the decision boundary (Boser et al., 1992; Cortes & Vapnik, 1995). Due to its superior performance, SVM has widespread application demands in many fields, such as intelligent monitoring, human-computer interaction and virtual reality (Yang & Gao, 2020). Anwar et al. (2019) proved that SVM cubic kernels with MFCC achieved approximately 96.7% accuracy for amateur drone detection. The SVM trained model can be considered hyperplane, with samples separated into two classes, which can divide data correctly and spaced at the largest interval with each sample (Rai et al., 2016). In this research, there were three SVM models that used three kernel functions to classify beehive sounds.

The formula for the linear kernel:

$$K\left(x_i, x_j\right) = x_i^T x_j$$

1

The formula for the polynomial kernel:

$$K\left(x_i, x_j\right) = \left(x_i^T x_j\right)^d$$

2

where d is the power parameter.

The formula for the RBF kernel:

$$K\left(x_i, x_j\right) = e^{-\gamma \| x_i - x_j \|^2}$$

3

where $\|x_i - x_j\|^2$ means the ethic distance between the two feature vectors, and γ is the width parameter of the RBF.

To improve the accuracy of classification of SVM, we optimized parameters C and γ in the kernel function. We set some values for C (0.1, 1, 10, 100, 1000) and γ (0.01, 0.1, 1). Then, we tried to find the optimal parameters through different kernel models. Finally, we chose a group of parameters (C and γ) with the highest accuracy, and they were used in SVM models.

In this study, SVM, RF and KNN were used to separately establish a classification model. Then, we compared the accuracy between the SVM, RF and KNN models. Finally, we chose the best model and used the data extracted at the beginning to test the accuracy of identification under different treatments.

Results

Data classification

After the preprocessing of data, a total of 2 160 audio sample data points were used in data mining and ML. In the data, each 720 audio samples was related to PS, SA or SE. Then, a total of 5 760 audio samples, which were collected from the two orderly feeding sessions, were used for sound recognition. Each feeding session contained 2 880 data samples, and they were evenly distributed in four stages.

Feature Extraction

To determine the number of features required for ML, recursive feature elimination (RFE) was performed through the RF model (Granitto et al., 2006). In RFE, models are built over and over again to pick the best features. Then, the process is repeated with the remaining features until all the features are traversed. The relationship between the number of features and the cross-validation score is shown in Fig. 3(b). In the early stages, the score rose dramatically as the number of features used increased. Then, the cross-validation score growth slows after the number of features used exceeds 10 and peaks when the number

of features surge to 20. After that, it slowly declines with fluctuations. Therefore, to build models with higher accuracy, the top 20 important features were used in this study. Figure 3(a) compares the importance of the variable and sorted by numerical size. The data in Fig. 3 suggest what features we should use to build ML models.

Precision Comparison Of Models

After the screening of characteristics, we built the models with KNN, RF and SVM (which used three different kernel functions) through the audio data processed by RFE. In addition, we tested different parameters to obtain a better model. The results of the correlational analysis are summarized in Table 1. The model trained by KNN, RF and SVM (using the RBF kernel function) had greater accuracy than the remaining models. One of the most accurate models is SVM (RBF). When the box constraint C was 1 and the value of γ was 0.1, the accuracy, macroaverage and weighted average were 95%. When the number of neighbors K is 4, the accuracy of KNN is 91%, the macroaverage is 94% and the weighted average is 94%. When the number of estimators is 91, the accuracy, macroaverage and weighted average are 91%. For the identification of different syrup types, all models have the highest accuracy for PS. The accuracy of each model is more than 93%. The second easily identifiable treatment is SE, and the last is SA. According to the results of Table 1, we chose the model with SVM (RBF) to detect the identification accuracy of different treatments.

Table 1
Classification result of different model

Treatment of Syrup	KNN	RF	SVM (Linear)	SVM (Quadratic)	SVM (RBF)
SA	91%	88%	94%	94%	**94%**
SE	96%	93%	96%	96%	**96%**
PS	94%	92%	95%	95%	**96%**
Accuracy	91%	91%	95%	95%	**95%**
Macro average	94%	91%	95%	95%	**95%**
Weighted average	94%	91%	95%	95%	**95%**

Prediction Of Different Syrups

In the next stage, we compared the accuracy between one type of syrup and the same type of syrup, which were affected by other chemicals. This calculation is done through the SVM model. The treatment that was not affected by other chemical compounds was named CK, and another treatment was named T.

Figure 4 shows the recall rate and accuracy of PS that is first fed in two orderly feeding processes. There were 712 and 713 data points in 720 audio samples of the first and the second feeding processes correctly identified, respectively (Fig. 4a). In addition, 1187 and 1097 audio samples were identified as PS. The error rates are 40% and 35%, respectively (Fig. 4(b)).

Figure 5(a) compares the recall data of SA in different situations. In cases not affected by ethyl acetate, the model with SVM (RBF) successfully identified 641 samples from 720, with 79 errors. For the treatment that had been fed SE, the correct number dropped to 216. Therefore, the result of the chi-square test reveals that there was a significant difference in accuracy between the two sets of data. The result of the identification rate of SE influenced by acetone is displayed in Fig. 5(b). For the treatment that had been fed by SA, the model correctly identified 267 samples from 720. This figure is smaller than the number, which is 576, of the group without being impacted by acetone. Judging from the results of significant differences, acetone also had a noticeable influence on the identification of ethyl acetate.

The result of the identification of PS is similar to those influenced by chemical compounds. It can be observed from Fig. 5(c) and Fig. 5(d) that compared with the correct rate of the identification of PS, the recall rate became lower when affected by a chemical compound regardless of whether it was acetone or ethyl acetate. For ethyl acetate, the number of successfully distinguished samples decreased from 712 to 317 in 720. The same type of data for acetone dropped from 713 to 382. These data show that the recall rate in identifying PS is significantly affected by compounds.

The precision results of SA and SE are shown in Fig. 6(a) and Fig. 6(b). In the absence of the effects of ethyl acetate, the model treated 763 audio samples as acetone with a correct rate of approximately 84%. In the treatment that had been fed SE, the number dropped to 281 with a correct rate of approximately 77%. For the treatment that had been fed by SA, the model identifies 290 audio samples, as SE with 267 was true. This figure is smaller than the number, which is 744, of the group without being impacted by acetone. From the results of significant differences, acetone and ethyl acetate had a significant effect on the precision of identification of each other.

Figure 6(c) and Figure 6(d) show that chemical compounds have a significant effect on the identification of syrup. There were 1061 and 1097 audio samples that were influenced or not by acetone and were identified as syrup with precisions of 36% and 65%, respectively. After being affected by ethyl acetate, the number of samples considered PS decreased from 1187 to 1174. In addition, accuracy dropped from 60–27%. These data illustrate that there are significant differences in whether the accuracy of PS identification is affected by chemical compounds.

Discussion

There are various messages regarding the sound of honeybee swarms. From the third century BC to the 19th century AD, many researchers have observed that bees make special sounds in some cases (Terenzi et al., 2020). Since the second half of the 20th century, due to the development of modern electronics, scientists have been able to record and analyse sounds in beehives. Especially in recent studies, ML has

been used in the sound analysis (Ribeiro et al., 2021). Therefore, the beehive sound characteristics used in ML are becoming increasingly important. In this study, by the method of data mining, we found that the low-level signal features and MFCCs are both integral parts of ML. MFCC features still showed more importance than the low-level signal features in the top 20 characteristics calculated by RF. By selecting the right data, we were able to build more accurate models. In this study, the ML model based on the SVM (RBF) model had the highest accuracy, which was approximately 95% (Table 1).

Research by Sharif (2020) has shown that the sound of beehives can be influenced by volatile chemical compounds in the air, such as acetone, ethyl ether, glutaric dialdehyde and trichloromethane. In addition, these changes can be identified by the ML model. In this experiment, the results of the syrup prediction were compared at the first phase of the two orderly feeding sessions. According to the recall results, the vast majority of PS samples were identified, and there was no significant difference between the two sets of data (Fig. 4(a)). This conclusion indicates that two orderly feeding sessions can be considered independent and have no influence on each other. In the next stages, Fig. 5 shows that both SA and SE significantly reduced the recall rate of another treatment and PS. In each continuous feeding experiment, the collected syrup was not removed from the frames, and the interval time between each stage was short. Chemical compounds can be thought to remain in beehives. Therefore, we speculate that both acetone and ethyl acetate that remained in the beehive could cause errors of identification about PS, SA or, SE.

The results based on the precision of the forecast show that the ML model gives the most predictions as PS. Without being influenced by chemical compounds, over 60% of these samples are predicted correctly. After feeding SA and SE, the ratio was reduced by half. For SA, SE samples that are not affected, the forecast given by the model is slightly higher than the actual sample size. And after feeding SE or SA, the number computed by the model is much lower than the actual sample size. However, there was no significant reduction in accuracy. The above results show that many SA and SE samples are judged as PS by the model, regardless of whether they are affected by chemical compounds. There was no significant change in the predicted number of SA and SE samples with no influence of chemical compounds in beehives. At the same time, in the case of ethyl acetate or acetone contained in beehives, the number of predictions made by the ML model is significantly reduced.

Combining the results of recall and precision, we found that the ML model can accurately identify PS. However, in the calculation, some SA and SE processes are predicted as PS. This may be due to honeybees collecting honey from other sources and transmitting signals to beehives. When these signals are transmitted to the beehive, the characteristics of the sound in that segment are classified as PS instead of SA or SE. In the case of residual chemicals in beehives, the ML model tended to judge fewer samples as SA or SE. As a result, a large number of SA or SE samples are judged to be other treatments, resulting in a significant decrease in the recall rate.

This is because when honeybees contain acetone or ethyl acetate in their beehives, the smell of the compound spreads in closed beehives (Zhao et al.,2021). Then, we made honeybees collect syrup

containing another compound. If the sound of beehives is mainly affected by smells, the accuracy of the same model can be greatly affected compared to identifying a chemical compound without interference. This result illustrates that chemical compounds that remain in the beehive affect the sound of the beehive. However, new chemicals entering the beehive can still be identified by changes in beehive sound. As a result, chemicals alter the sound of beehives not only through smell in the air but also probably through the method by which honeybees transmit food signals to the colony.

Honeybees produce different beehive sounds when stimulated by different chemical compounds. These differences make it possible to monitor contaminants by identifying changes in the sound of beehives. This also means that the sound of hives can be used not only to monitor pollutants in the atmosphere but also to monitor pollution in water and soil. It is worth noting that it is a great method to increase the number of experimental data to improve the accuracy of each ML model.

In this research, all samples were collected during sunny day, and under similar climatic conditions. Beehive sound intensity may be correlated with temperature in the hive, coupled with the weather (Imoize et al., 2020). Therefore, even if it is possible to reduce the accuracy of the model, it is necessary to add beehive sound data under different weather conditions in future studies. In the future, we plan to collect beehive sounds of different compounds in different weather conditions. Moreover, we intend to assess the effects of nonvolatile compounds or heavy metals on beehive sound.

Conclusion

In this study, we collected beehive sounds and extracted important features by data mining. Then, classification models used to identify acetone and ethyl acetate were established. The results of the ML models were used to classify compounds. The following conclusions can be obtained. Honeybees are likely to report chemical compounds in beehives through colony sound. In addition, compounds outside the beehive can also be monitored by beehive sound.

Declarations

Acknowledgements

The authors appreciate the assistance of Sericultural & Apicultural Research Institute, Yunnan Academy of Agricultural Sciences for their help in the experiment process.

Funding

This work was financially supported by the Hefei Institutes of Physical Science, the Chinese Academy of Sciences.

Author contribution

BY: conceptualization, methodology, investigation, data processing, visualization, writing; XH: methodology, investigation; MZS: methodology, review, editing; XJ: suggestion, review; ND: suggestion, review; FL: supervision, suggestion, review, editing.

Data availability

Not applicable.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Anitha P, Kaarthick B (2021) Oppositional based Laplacian grey wolf optimization algorithm with SVM for data mining in intrusion detection system. J Ambient Intell Humaniz Comput 12(3):3589–3600
2. Anwar MZ, Kaleem Z, Jamalipour A (2019) Machine learning inspired sound-based amateur drone detection for public safety applications. IEEE Trans Veh Technol 68(3):2526–2534
3. Balayiannis G, Balayiannis P (2008) Bee honey as an environmental bioindicator of pesticides' occurrence in six agricultural areas of Greece. Arch Environ Contam Toxicol 55(3):462–470
4. Barati E, Saraee MH, Mohammadi A, Adibi N, Ahmadzadeh MR (2011) A survey on utilization of data mining approaches for dermatological (skin) diseases prediction. J Sel Areas Health Inf (JSHI) 2(3):1–11
5. Bartling MT, Thümecke S, Russert JH, Vilcinskas A, Lee KZ (2021) Exposure to low doses of pesticides induces an immune response and the production of nitric oxide in honeybees. Sci Rep 11(1):1–11
6. Boser BE, Guyon IM, Vapnik VN (1992), July A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152)
7. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
8. Bromenshenk JJ, Henderson CB, Seccomb RA, Rice SD, Etter RT (2009) *U.S. Patent No. 7,549,907*. Washington, DC: U.S. Patent and Trademark Office
9. Calle ML, Urrea V (2011) Letter to the editor: stability of random forest importance measures. Brief Bioinform 12(1):86–89
10. Cejrowski T, Szymański J, Logofătu D (2020) Buzz-based recognition of the honeybee colony circadian rhythm. Comput Electron Agric 175:105586
11. Chauhan VK, Dahiya K, Sharma A (2019) Problem formulations and solvers in linear SVM: a review. Artif Intell Rev 52(2):803–855

12. Chen Y, Zhou L, Pei S, Yu Z, Chen Y, Liu X, Xiong N (2019) KNN-BLOCK DBSCAN: Fast clustering for large-scale data.IEEE Transactions on Systems, Man, and Cybernetics: Systems
13. Chen Z, Li J, Wei L, Xu W, Shi Y (2011) Multiple-kernel SVM based multiple-task oriented data mining system for gene expression data analysis. Expert Syst Appl 38(10):12151–12159
14. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366
15. Cochard P, Laurie M, Veyrand B, Le Bizec B, Poirot B, Marchand P (2021) PAH7 concentration reflects anthropization: A study using environmental biomonitoring with honeybees. Sci Total Environ 751:141831
16. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
17. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
18. Dicks L, Breeze T, Ngo H, Senapathi D, An J, Aizen M, Potts S (2020) A global assessment of drivers and risks associated with pollinator decline
19. Granitto PM, Furlanello C, Biasioli F, Gasperi F (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr Intell Lab Syst 83(2):83–90
20. Han BA, O'Regan SM, Schmidt P, Drake JM (2020) Integrating data mining and transmission theory in the ecology of infectious diseases. Ecol Lett 23(8):1178–1188
21. Imoize AL, Odeyemi SD, Adebisi JA (2020) Development of a Low-Cost Wireless Bee-Hive Temperature and Sound Monitoring System. Indonesian J Electr Eng Inf (IJEEI) 8(3):476–485
22. Kawakita S, Ichikawa K, Sakamoto F, Moriya K (2019) Sound recordings of Apis cerana japonica colonies over 24 h reveal unique daily hissing patterns. Apidologie 50(2):204–214
23. Li Y, Zhang J, Li T, Liu H, Li J, Wang Y (2017) Geographical traceability of wild Boletus edulis based on data fusion of FT-MIR and ICP-AES coupled with data mining methods (SVM). Spectrochim Acta Part A Mol Biomol Spectrosc 177:20–27
24. Lin KWE, Feng T, Agus N, Therefore C, Lui S (2014), December Modelling mutual information between voiceprint and optimal number of mel-frequency cepstral coefficients in voice discrimination. In *2014 13th International Conference on Machine Learning and Applications* (pp. 15-20). IEEE
25. Liu Z, Li S, Hao J, Hu J, Pan M (2021) An Efficient and Fast Model Reduced Kernel KNN for Human Activity Recognition. *Journal of Advanced Transportation*, *2021*
26. Logan B (2000) Mel frequency cepstral coefficients for music modeling. In In International Symposium on Music Information Retrieval
27. Luo QH, Gao J, Guo Y, Liu C, Ma YZ, Zhou ZY, Diao QY (2021) Effects of a commercially formulated glyphosate solutions at recommended concentrations on honeybee (Apis mellifera L.) behaviours. Sci Rep 11(1):1–8
28. Mahapatra B, Dhal NK, Dash AK, Panda BP, Panigrahi KCS, Pradhan A (2019) Perspective of mitigating atmospheric heavy metal pollution: using mosses as biomonitoring and indicator

organism. Environ Sci Pollut Res 26(29):29620–29638

29. Mei Q, Gül M, Boay M (2019) Indirect health monitoring of bridges using Mel-frequency cepstral coefficients and principal component analysis. Mech Syst Signal Process 119:523–546
30. Mistry P, Neagu D, Trundle PR, Vessey JD (2016) Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. Soft Comput 20(8):2967–2979
31. Müller M, Kurth F, Clausen M (2005), September Audio Matching via Chroma-Based Statistical Features. In ISMIR (Vol. 2005, p. 6)
32. Nolasco I, Benetos E (2018) To bee or not to bee: Investigating machine learning approaches for beehive sound recognition. *arXiv preprint arXiv:1811.06016*
33. Ollerton J, Winfree R, Tarrant S (2011) How many flowering plants are pollinated by animals? Oikos 120(3):321–326
34. Pérez N, Jesús F, Pérez C, Niell S, Draper A, Obrusnik N, Monzón P (2016) Continuous monitoring of beehives' sound for environmental pollution control. Ecol Eng 90:326–330
35. Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE (2010) Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol 25(6):345–353
36. Rai P, Golchha V, Srivastava A, Vyas G, Mishra S (2016), August An automatic classification of bird species using audio feature extraction and support vector machines. In *2016 International Conference on Inventive Computation Technologies (ICICT)* (Vol. 1, pp. 1-5). IEEE
37. Ribeiro AP, da Silva NFF, Mesquita FN, Araújo PCS, Rosa TC, Mesquita-Neto JN (2021) Machine learning approach for automatic recognition of tomato-pollinating bees based on their buzzing-sounds. PLoS Comput Biol 17(9):e1009426–e1009426
38. Robles-Guerrero A, Saucedo-Anaya T, González-Ramírez E, De la Rosa-Vargas JI (2019) Analysis of a multiclass classification problem by lasso logistic regression and singular value decomposition to identify sound patterns in queenless bee colonies. Comput Electron Agric 159:69–74
39. Robertson HM, Wanner KW (2006) The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family. Genome Res 16(11):1395–1403
40. Sari MF, Esen F, Tasdemir Y (2021) Levels of polychlorinated biphenyls (PCBs) in honeybees and bee products and their evaluation with ambient air concentrations. Atmos Environ 244:117903
41. Sharif MZ, Wario F, Di N, Xue R, Liu F (2020) Soundscape Indices: New Features for Classifying Beehive Audio Samples. Sociobiology 67(4):566–571
42. Smith KE, Weis D, Amini M, Shiel AE, Lai VWM, Gordon K (2019) Honey as a biomonitor for a changing world. Nat Sustain 2(3):223–232
43. Tan PN, Steinbach M, Kumar V (2006) Introduction to Data Mining
44. Tashakkori R, Hamza AS, Crawford MB (2021) Beemon: An IoT-based beehive monitoring system. Comput Electron Agric 190:106427
45. Team RC (2013) R: A language and environment for statistical computing

46. Terenzi A, Cecchi S, Spinsante S (2020) On the importance of the sound emitted by honey bee hives. Veterinary Sci 7(4):168
47. Themudo GE, Rey-Iglesia A, Tascón LR, Jensen AB, da Fonseca RR, Campos PF (2020) Declining genetic diversity of European honeybees along the twentieth century. Sci Rep 10(1):1–12
48. Vizcaino MI, Guo X, Crawford JM (2014) Merging chemical ecology with bacterial genome mining for secondary metabolite discovery. J Ind Microbiol Biotechnol 41(2):285–299
49. Waltham NJ, Teasdale PR, Connolly RM (2013) Use of flathead mullet (Mugil cephalus) in coastal biomonitor studies: review and recommendations for future studies. Mar Pollut Bull 69(1–2):195–205
50. Wang CX, Li YY, Xu SQ (2010) Biological monitoring and its application in environmental monitoring. Asian J Ecotoxicol 5(5):628–638
51. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
52. Xia X, Togneri R, Sohel F, Huang D (2018) Random forest classification based acoustic event detection utilizing contextual-information and bottleneck features. Pattern Recogn 81:1–13
53. Xu DM, Liu WL, Liu WP (2009) Research advances in toxicological effects of external pollutants on earthworms. A-sian J Ecotoxicol 4(1):21–27
54. Xu Y, Zhang Q, Wang L (2018) Metric forests based on Gaussian mixture model for visual image classification. Soft Comput 22(2):499–509
55. Yang J, Gao H (2020) Cultural emperor penguin optimizer and its application for face recognition. *Mathematical Problems in Engineering*, *2020*
56. Yasodha S, Prakash PS (2012), March Data mining classification technique for talent management using SVM. In 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET) (pp. 959-963). IEEE
57. Zacepins A, Brusbardis V, Meitalovs J, Stalidzans E (2015) Challenges in the development of Precision Beekeeping. Biosyst Eng 130:60–71
58. Zhao Y, Deng G, Zhang L, Di N, Jiang X, Li Z (2021) Based investigate of beehive sound to detect air pollutants by machine learning. Ecol Inf 61:101246

Figures

Figure 1

Experimental location and design.

The experimental conditions were consistent, and the experiment was conducted on sunny days with similar temperatures and humidity in August 2021.

Figure 2

Procedure of a single syrup (a) and the whole experiment (b).

Figure 3

The importance order of the top 30 feature sets (a) and the curve made by RFE (b).

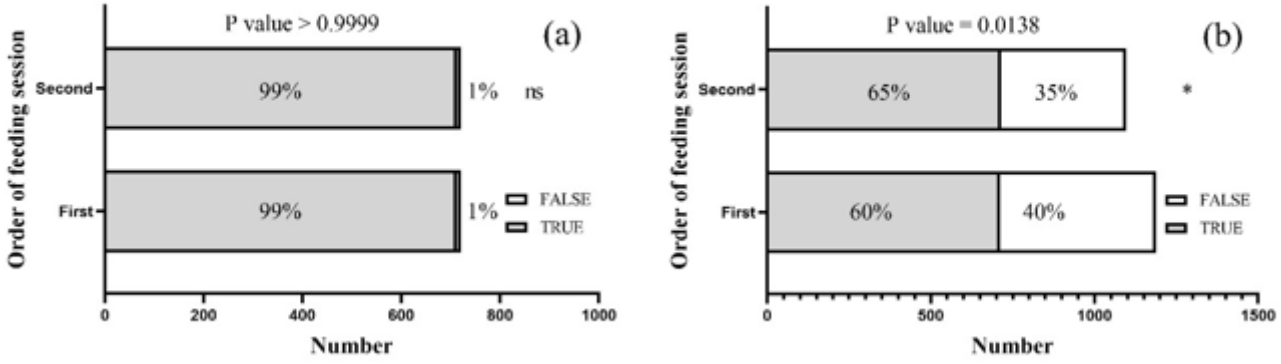

Figure 4

Comparison of recall (a) and precision (b) between two orderly feeding sessions (ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p<0.001$, ****: $p < 0.0001$).

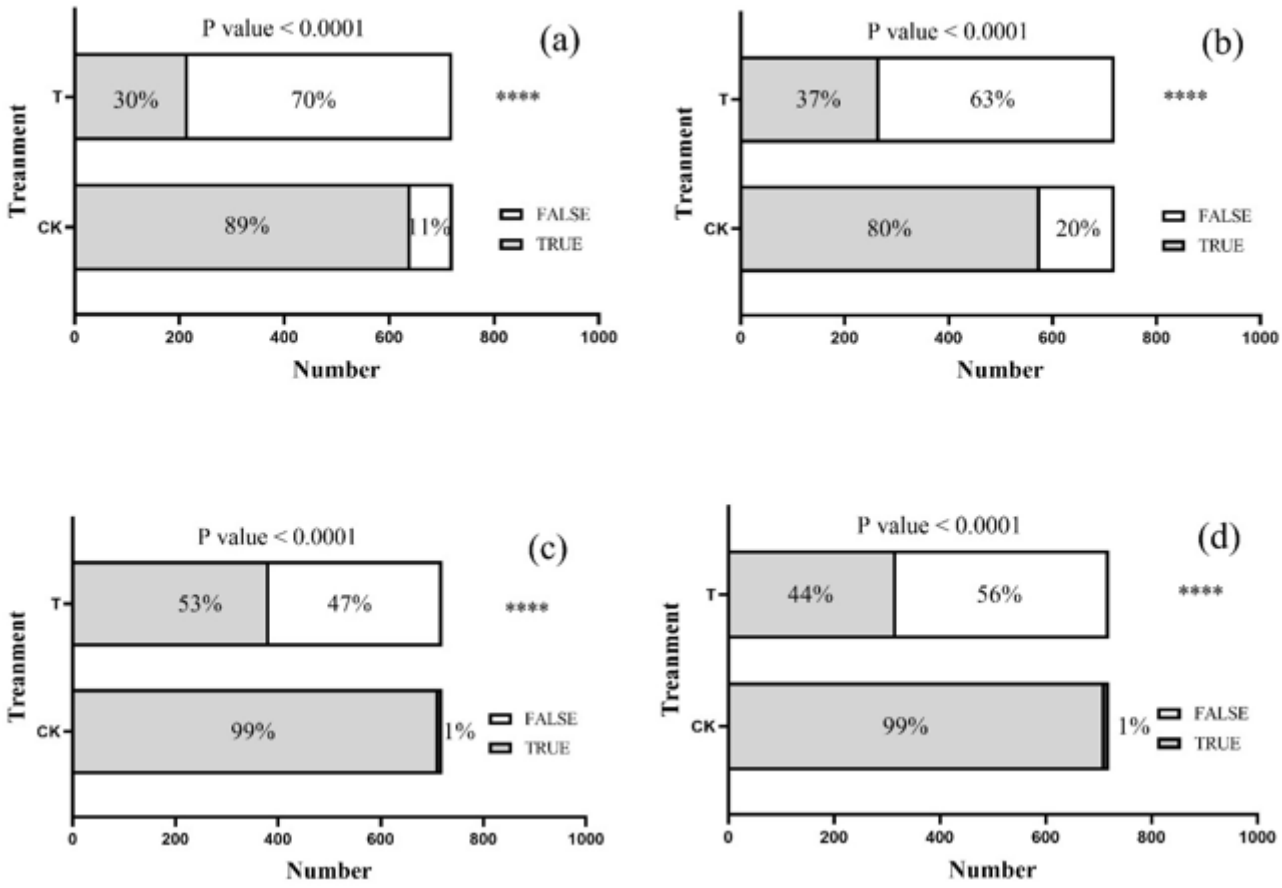

Figure 5

Comparison of recall rate between syrup unaffected by chemical compound (CK) and similar syrup affected by compound (T): (a) ethyl acetate influenced acetone, (b) acetone influenced ethyl acetate, (c) acetone influenced blank syrup, (d) ethyl acetate influenced blank syrup (ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p<0.001$, ****: $p < 0.0001$).

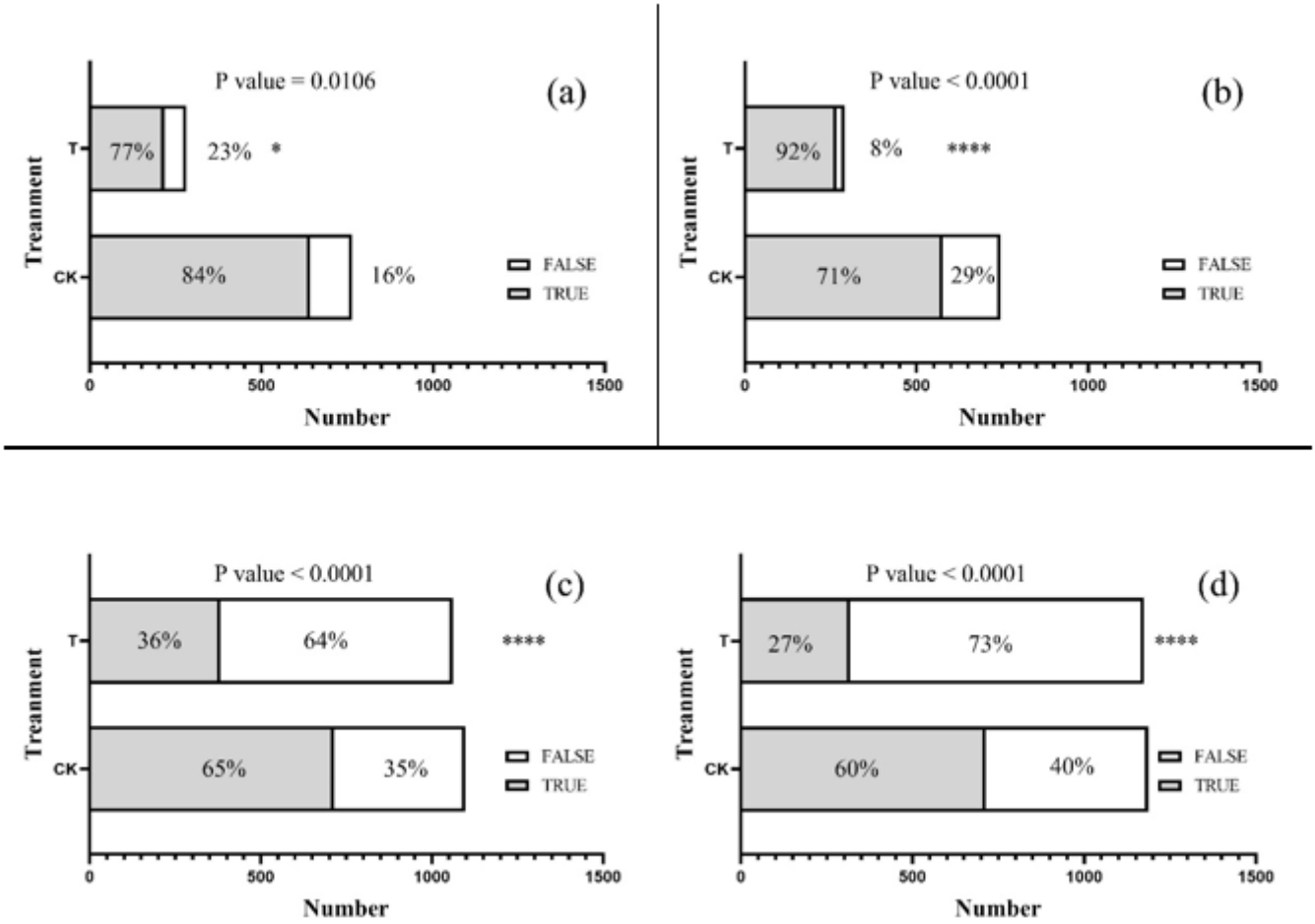

Figure 6

Comparison of precision between syrup unaffected by chemical compound (CK) and similar syrup affected by compound (T): (a) ethyl acetate influenced acetone, (b) acetone influenced ethyl acetate, (c) acetone influenced blank syrup, (d) ethyl acetate influenced blank syrup (ns: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p<0.001$, ****: $p < 0.0001$).