# Machine Learning Advances in Beekeeping

Sonja Dimitrijević*, Nikola Zogović*
* University of Belgrade, Mihailo Pupin Institute, Belgrade, Serbia
sonja.dimitrijevic@pupin.rs; nikola.zogovic@pupin.rs

***Abstract* — Digitalization has brought IoT and big beehive data in beekeeping. Consequently, the need for advanced analysis of beehive data and other bee-related data has led to machine learning. The objective of the paper is to reveal research tendencies, trends and gaps in recent literature on machine learning advances in beekeeping. Therefore, the paper reviews articles that have been published in the last two years in good-quality journals. More specifically, the review analyzes and discusses applications, (sub)types and algorithms of machine learning, including data sets used, that have been reported in selected articles. The results have shown that the most common application cases are related to bee products quality assessment/authentication and identification/prediction of beehive conditions. To this end, supervised learning, more specifically, classification, has been predominantly used. The most often used algorithms are Random Forest, Support Vector Machine and Convolutional Neural Network. Moreover, a variety of data has served as input to the models including images, sound, data from beehive sensors, meteorological data, spectroscopy data on honey samples, etc. Machine learning in beekeeping is still in its early years. However, many opportunities have already been identified and promising research directions are just opening out.**

## I. INTRODUCTION

Bees are among the main pollinators in diverse ecosystems, and consequently significant contributors to world nutritional security. In addition to its importance in food production, animal pollination is economically, socially, and culturally relevant due to its impact on the production of biofuels, fibers, medicines, and building materials [1]. Moreover, honey bees produce honey, royal jelly, propolis, pollen, beeswax and venom that have been well known for nutritional or health benefits for centuries [2]. Thanks to numerous active biological features (such as antioxidant, antimicrobial, antiviral, anti-inflammatory, and antitumoral), bee products are used for different traditional and complementary therapies, also known as apitherapy [3].

Beekeeping has a long history dating back to the Neolithic period. It refers to the continuum from a hobby to a full-time professional occupation. Although professional beekeepers are by far in the minority, their production is substantial [2]. According to Bellos et al. [4], beekeepers face the following major challenges nowadays:

- beehive theft,
- bee health problems and particularly colony mortality risk due to illnesses and varroa mite infection,
- extreme weather conditions,
- animal attacks,
- colony losses due to unfortunate farming practices, and
- production of high-quality products following production standards.

Modern threats to honey bee populations such as pesticides, parasites, predators and diseases have increased awareness of the economic and social importance of beekeeping [5].

Information and communication technologies have brought many new opportunities in agriculture, including beekeeping. The development of sensors has enabled efficient and economically viable bee-related data collection and transmission. Therefore, the Internet of Things (IoT) has come into play and consequently, the need for advanced data analysis has emerged [2]. Artificial intelligence (AI), particularly machine learning, has offered ways how to tackle the problems of big beehive data analysis and advanced analysis of many other bee-related data such as data on the toxicity of pesticides for honey bees [6] or spectroscopy data on honey samples for honey authentication [7].

Machine learning is an essential part of AI that in essence enables computers to learn without explicit programming. There are 3 broad types of machine learning processes: supervised, unsupervised and reinforcement learning. Supervised learning finds the relations between inputs and outputs (datasets with labels) while training the system. The main subtypes of supervised learning are classification and regression. In unsupervised learning, there is no output mapped to the inputs; the model extracts the relations from the data. It is primarily used for grouping similar patterns into clusters (clustering) and a low-dimensional space (dimensionality reduction). Finally, reinforcement learning algorithms constantly learn by interacting with the environment in order to take certain actions [8].

A branch of machine learning that has gotten much hype in many research fields is deep learning. Broadly speaking, deep learning uses neural networks with many layers, which have enabled the extraction of high-level features from the data, working with or without labels and training models for the fulfillment of multiple objectives [8].

In the last couple of decades, there has been an increase in research papers whose topics are related to beehive monitoring, i.e., smart beehives. Their focus is on bee-related data collecting and transferring processes, as well as on supporting technology. On the other hand, an increase in research papers with a focus on bee-related data analysis, including the use of machine learning to this end, has been evident in the last few years. However, it is not clear for what specific purposes machine learning has been applied in research on beekeeping, how it has been

applied and what data have been used. To the best of the authors' knowledge, no review addresses such questions.

Therefore, this review analyzes and discusses applications, (sub)types and algorithms of machine learning, including data sets used, that have been reported in selected articles. Moreover, it reveals research tendencies, trends and gaps in recent good-quality literature on machine learning in beekeeping. Therefore, the review is of benefit for researchers in the fields of both apiculture and AI. It may indicate new research perspectives to researchers and help them make more informed decisions on how to direct their research.

## II. Research Questions

This paper reviews recent advances in beekeeping as reported in literature. The objective of the paper is to summarize and discuss how machine learning has been applied in beekeeping. Therefore, the following review questions have been defined:

RQ1: What applications of machine learning have been reported in the reviewed literature?

RQ2: What machine learning (sub)types and algorithms have been reported in the reviewed literature?

RQ3: What data types/sets have been used in the reported applications of machine learning?

The review focuses on algorithms on which machine learning models are based. In other words, algorithms applied in data preprocessing including dimensionality reduction and feature selection have not been considered.

To the best of the authors' knowledge, there is no other review paper with a specific focus on recent machine learning advances in beekeeping. A couple of other reviews on technological advances in beekeeping have a broader or only a tangentially related scope without a focus on machine learning. For instance, Hadjur et al. [2] review precision beekeeping systems and services, whereas Sharif et al. [9] review applications of big beehive data in climate-smart architecture.

## III. Methodology

The review process has involved the following phases:

- search for papers
- selection of relevant papers
- data extraction
- data analysis and discussion

A search for papers has been conducted using Google Scholar due to the wide search capabilities. The used search term is "machine learning beekeeping". Since the review targets recent studies, only papers published in 2020 and after have been considered. The search has returned over 700 results that have been examined according to the inclusion criteria in the second phase of the review.

The following inclusion criteria have been defined:

- a study reported in the article applies machine learning to bee-related data
- the article is published in a journal with an impact factor
- the article is in English

First, a search result has been checked according to the last two criteria. Then, the remaining articles have been read to examine compliance with the first criterion. At the end of this phase, 43 papers have been selected for data extraction.

In the phase of data extraction, a spreadsheet table has been used. This table has columns such as reference, machine learning application, application domain, machine learning (sub)type, algorithms, data sets and comments. The columns intended for the extraction of data on machine learning applications and domains address RQ1. Machine learning (sub)types and algorithms have been extracted to address RQ2. Data sets in terms of main data groups/types have been extracted to address RQ3. Comments are used to provide the context.

In the final phase, the data extracted from selected articles have been analyzed, summarized and discussed according to the research questions.

## IV. Results

The review results are presented according to the research questions defined in section II.

### *A. Applications of machine learning in beekeeping*

Regarding RQ1 (What applications of machine learning have been reported in the reviewed literature?), many different machine learning applications have been reported in the reviewed articles. These applications can be categorized into several domains/categories. Table 1 shows the distribution of studies according to the most common application domains.

10 out of 43 studies have applied machine learning for bee products quality assessment and/or bee products authentication (e.g., [7], [10], [11]). Honey is especially vulnerable to adulteration with cheaper sugars and/or syrups. Moreover, due to the premium price of mono-floral and mono-geographical products, false origin labeling or misdescription is not a rare fraud practice [10]. Therefore, these studies have mostly tackled such problems thanks to data that originate from novel approaches based on vibrational spectroscopy and laser-induced breakdown spectroscopy.

The second most-often reported application domain is pest or disease presence identification (5 out of 43 studies) (e.g. [12], [13]), which can be seen as a subdomain of identification of honey bee colony health status (3 out of 43 studies) (e.g., [14], [15]).

Analysis of honey bee behavior has been also the subject of 5 studies in total. This diverse domain has involved analyses of bee behavior such as foraging behavior (e.g., [16]), behavior in queenless bee colonies [17], etc. As in the previous two application domains, analysis for this purpose is mostly based on beehive monitoring data.

Analysis of climate change impact on the distribution of bees or pests (e.g., [18]), as well as, analysis of plant protection products (PPPs) toxicity (e.g., [6]) are application domains of machine learning in 3 studies each. PPPs and climate change are both among modern threats to bee populations. Accordingly, studies addressing such major problems in the present and future of beekeeping are expected.

Other application domains such as swarming in honey bee colonies (e.g., [19]) or bee (sub)species identification [20] have been studied in less than 3 studies.

TABLE I.
DISTRIBUTION OF STUDIES ACCORDING TO THE MOST COMMON APPLICATION DOMAINS

| Application domain | Number of studies | Percentage of studies |
|---|---|---|
| bee products quality assessment and/or authentication | 10 | 23.3% |
| pest or disease presence identification | 5 | 11.6% |
| analysis of honey bee behavior | 5 | 11.6% |
| identification of honey bee colony health status | 3 | 7% |
| analysis of climate change impact on the distribution of bees or pests | 3 | 7% |
| analysis of plant protection products (PPPs) toxicity | 3 | 7% |

### *B. Machine learning (sub)types and algorithms in beekeeping*

When it comes to RQ2 (What machine learning (sub)types and algorithms have been reported in the reviewed literature?), supervised learning, more specifically, classification has been predominantly applied in the reviewed studies. Table 2 shows that another subtype of supervised learning, regression, has been applied in 6 studies. Together, classification and regression have been applied in 3 studies. For instance, both classification and regression models have been used for honey harvest prediction [21].

Unsupervised learning, more precisely clustering, has been applied only in one study to identify seasonal patterns of honey bees. The classification conducted in the second phase of the same study has been based on the identified clusters [22].

TABLE II.
DISTRIBUTION OF STUDIES ACCORDING TO THE (SUB)TYPES OF MACHINE LEARNING

| Machine learning type | Machine learning subtype | Number of studies | Percentage of studies |
|---|---|---|---|
| supervised learning | classification | 40 | 93% |
| | regression | 6 | 14% |
| unsupervised learning | clustering | 1 | 2.3% |
| reinforcement learning | object detection | 1 | 2.3% |

Object detection, more specifically, the detection of bees from video streams in real-time, has been applied in two studies. Additionally, classification has been used in both studies for the recognition of pollen and non-pollen bearing honey bees [16], i.e., bees infected or non-infected with Varroa destructor [23]. The authors of the second study claim that their approach relies on reinforcement learning.

Specific machine learning algorithms that have been used in more than 4 studies are:

- Random Forest (15 studies; 14.4%),
- Support Vector Machine (SVM) (11 studies; 10.6%),
- Convolutional Neural Network (CNN) (8 studies; 7.7%),
- K-Nearest Neighbors (KNN) (6 studies; 5.8%),
- Decision Trees (4 studies; 3.8%),
- Logistic Regression (4 studies; 3.8%), and
- Soft Independent Modeling of Class Analogy (SIMCA) (4 studies; 3.8%).

Most of the studies have applied more than one algorithm mostly for comparison purposes. In addition to CNN, deep learning has been implemented in a few more studies using:

- Long Short-Term Memory (LSTM) neural network (2 studies),
- Feedforward Neural Network (FNN) (2 studies),
- Gated Recurrent Units (GRU) neural network (1 study),
- Recurrent Neural Network (RNN) (1 study),
- Graph Attention Convolutional Neural Network (GACNN) (1 study), and
- Deep Neural Network (DNN) classifier (1 study).

LSTM and GRU have been used for regression problems. In addition, object detection has been implemented using a pair of RNN (parent) and CNN (child) in one of the studies [23], as well as YOLOv3-tiny (CNN-based object detector and classifier) in another study [16]. Other deep learning models have been used for classification problems.

### *C. Bee-related data sets/types used in machine learning*

As for RQ3 (What data types/sets have been used in the reported applications of machine learning?), various data have been used in the reviewed machine learning models. The data include:

- images (e.g., comb cells images, honey bee images for different conditions),
- sound (e.g., buzz sound),
- data from sensors inside (e.g., internal temperature, internal humidity, etc.) and outside a hive (e.g., external temperature, external humidity, etc.),
- meteorological data (e.g., temperature, rainfall, solar exposure, etc.),
- bioclimatic data,
- spectroscopy data and physicochemical parameters of bee product samples,
- plant-derived DNA sequences from honey samples
- molecular descriptors of PPPs,
- etc.

Spectroscopy data and physicochemical parameters of bee product samples have been used in the studies on bee products quality assessment and/or authentication (10 studies).

The studies on pest or disease presence identification have used data from gas sensors with or without infestation assessment (3 studies), microscopic images (1

study), and images/video frames from video streams (1 study).

On the one hand, some studies on the analysis of honey bee behavior have used specifically bee colony sound (1 study) or images/video frames from video streams (1 study). On the other hand, some studies addressing this application domain have combined different data types such as images/video frames from video streams, sensor data and meteorological data (1 study), as well as honey bee images (in a hive and et entrance) and some genetic data (1 study).

Likewise, one study on the identification of honey bee colony health status has used honey bee images for different conditions. The other two studies have combined different data types such as sensor data, behave mass and meteorological data.

Furthermore, the studies on the analysis of climate change impact on the distribution of bees or pests have minimally used environmental/bioclimatic data and data on species' occurrence or bee colonies inspections data.

In addition, the studies on the analysis of plant protection products (PPPs) toxicity have used molecular descriptors of pesticides and toxicity mechanisms as input data.

Many studies have collected their own data for the analysis. Some studies have used publicly available data.

## V. Discussion and Conclusion

Machine learning applications in studies on bee-related data analysis show that researchers try to address major risks and challenges in beekeeping. These risks and challenges include bee products quality assessment and/or authentication, bee colony health including the influence of PPPs on bee health, climate change impact on the distribution of bees or pests, prevention of colony losses, etc.

Studies on bee products quality assessment and/or authentication establish a research direction that shows some very promising results. These studies would benefit from a larger number and variety of samples of bee products though.

A variety of inputs and outputs in models on beehive monitoring data, despite a relatively small number of studies, suggest that a leading direction cannot yet be identified. A divided focus of some studies between data collection and data analysis, as well as alternative experiments in some of the studies, show that researchers parallelly consider what data to collect and how to get some valuable output. Yet, some of the studies use freely available data from open projects. In other words, their focus is fully on machine learning model development and/or improvement. However, open bee-related data sets are still scarce.

In addition, most studies compare different approaches/algorithms to find the best-suited ones for certain application cases. Nevertheless, they often lack data from multiple beehives, apiaries and locations. Consequently, the results of such comparisons are specific to a beehive/apiary/location and could be different when these limitations are overcome. Moreover, only a couple of studies have applied machine learning within an automated monitoring system.

Bee-related data analysis concerns different scientific fields such as agriculture, biology, ecology, chemistry, food science, etc., as the used data sets (or the titles of journals in which the articles were published) suggest. Consequently, a variety of application cases is not surprising.

The use of machine learning for bee-related data analysis is still in its early years though. Yet, many opportunities have already been identified and promising research directions are just opening out.

## Acknowledgment

The research described in this paper was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia [451-03-68/2022-14/200034].

## References

[1] P. G. Peixoto, H. L. Martins, B. C. Pinto, A. L. Franco, L. S. Amaral, and C. V. D. Castro, "The Significance of Pollination for Global Food Production and the Guarantee of Nutritional Security: A Literature Review", *Environ. Sci. Proc.*, vol. 15, no. 1, pp. 7, 2022.

[2] H. Hadjur, D. Ammar, and L. Lefèvre, "Toward an intelligent and efficient beehive: A survey of precision beekeeping systems and services", *Comput. Electron. Agric.*, no. *192*, pp. 106604, 2022.

[3] S. Kolayli, and M. Keskin, "Natural bee products and their apitherapeutic applications", *Stud. Nat. Prod. Chem.*, no. 66, pp. 175-196, 2020.

[4] C. V. Bellos, A. Fyraridis, G. S. Stergios, K. A. Stefanou, and S. Kontogiannis, "A Quality and disease control system for beekeeping", In *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)* (pp. 1-4), 2021.

[5] M. Roffet-Salque, M. Regert, R. P. Evershed, A. K. Outram, L. J. Cramp, O. Decavallas, ... and J. Zoughlami, "Widespread exploitation of the honeybee by early Neolithic farmers", *Nature*, vol. 527, no. 7577, pp. 226-230, 2015.

[6] F. Wang, J. F. Yang, M. Y. Wang, C. Y. Jia, X. X. Shi, G. F. Hao, and G. F. Yang, "Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction". *Sci. Bull.*, vol. 65, no. 14, pp. 1184-1191, 2020.

[7] J. P. Formosa, F. Lia, D. Mifsud, and C. Farrugia, "Application of ATR-FT-MIR for tracing the geographical origin of honey produced in the Maltese islands", *Foods*, vol. 9, no. 6, pp. 710, 2020.

[8] D. P. Kumar, T. Amgoth, and C. S. R. Annavarapu, "Machine learning algorithms for wireless sensor networks: A survey", *Inf Fusion*, vol. 49, pp. 1-25, 2019.

[9] M. Z. Sharif, N. Di, and F. Liu, "Monitoring honeybees (Apis spp.)(Hymenoptera: Apidae) in climate-smart agriculture: A review", *Appl. Entomol. Zool.*, pp. 1-15, 2021.

[10] T. Damiani, R. M. Alonso-Salces, I Aubone, V. Baeten, Q. Arnould, C. Dall'Asta, ... and J. A. Fernández Pierna, "Vibrational Spectroscopy Coupled to a Multivariate Analysis Tiered Approach for Argentinean Honey Provenance Confirmation". *Foods*, vol. 9, no. 10, pp. 1450, 2020

[11] N. A. Didaras, I. Kafantaris, T. G. Dimitriou, C. Mitsagga, K. Karatasou, I. Giavasis, ... and D. Mossialos, "Biological Properties of Bee Bread Collected from Apiaries Located across Greece", *Antibiotics*, vol. 10, no. 5, pp. 555, 2021.

[12] S. Dghim, C. M. Travieso-González, and R. Burget, "Analysis of the Nosema Cells Identification for Microscopic Images", *Sensors*, vol. 21, no. 9, pp. 3068, 2021.

[13] A. Szczurek, M. Maciejewska, B. Bąk, J. Wilk, J. Wilde, and M. Siuda, "Gas sensor array and classifiers as a means of varroosis detection", *Sensors*, vol. 20, no. 1, pp. 117, 2020.

[14] A. R. Braga, D. G. Gomes, R. Rogers, E. E., Hassler, B. M. Freitas, and J. A. Cazier, "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast

the health status of honey bee colonies". *Comput. Electron. Agric.*, vol. 169, pp. 105161, 2020b.

[15] 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. 89009-89019, 2021.

[16] T. N. Ngo, D. J. A. Rustia, E. C., Yang, and T. T. Lin, "Automated monitoring and analyses of honey bee pollen foraging behavior using a deep learning-based imaging system", *Comput. Electron. Agric.*, vol. 187, pp. 106239, 2021.

[17] B. M. Jones, V. D. Rao, T. Gernat, T. Jagla, A. C. Cash-Ahmed, B. E. Rubin, ... and G. E. Robinson, "Individual differences in honey bee behavior enabled by plasticity in brain gene regulatory networks", *Elife*, no. 9, pp. e62850, 2020.

[18] R. A. Giliba, I. H. Mpinga, S. A. Ndimuligo, and M. M. Mpanda, "Changing climate patterns risk the spread of Varroa destructor infestation of African honey bees in Tanzania", Ecol. Process., vol.9, no. 1, pp. 1-11, 2020.

[19] A. Zgank, "IoT-based bee swarm activity acoustic classification using deep neural networks", Sensors, vol. 21, no. 3, pp. 676, 2021.

[20] A. P. Ribeiro, N. F. F. da Silva, F. N. Mesquita, P. D. C. S. Araújo, T. C. Rosa, and J. N. Mesquita-Neto, "Machine learning approach for automatic recognition of tomato-pollinating bees based on their buzzing-sounds", PLoS Comput. Biol., vol. 17, no. 9, pp. e1009426, 2021.

[21] T. Campbell, K. W. Dixon, K. Dods, P. Fearns, and R. Handcock, "Machine learning regression model for predicting honey harvests", *Agriculture*, vol. 10, no. 4, pp. 118, 2020.

[22] A. R. Braga, D. G. Gomes, B. M. Freitas, and J. A. Cazier. "A cluster-classification method for accurate mining of seasonal honey bee patterns", *Ecol. Inform.*, vol. 59, pp. 101107, 2020a.

[23] D. Mrozek, R. Górny, A. Wachowicz, and B. Małysiak-Mrozek, "Edge-Based Detection of Varroosis in Beehives with IoT Devices with Embedded and TPU-Accelerated Machine Learning", *Appl. Sci.*, vol. 11, no. 22, pp. 11078, 2021.