Apidologie (2025) 56:5
© INRAE, DIB and Springer-Verlag France SAS, part of Springer Nature, 2025
https://doi.org/10.1007/s13592-024-01139-8

Original article

Explainable artificial intelligence for differentiating honey bee genotypes using morphometrics and SSR markers

Berkant İsmail Yildiz[1], Kemal Karabağ[1], Uğur Bilge[2], and Aziz Gül[3]

[1] Department of Agricultural Biotechnology, Faculty of Agriculture, Akdeniz University, 07058 Antalya, Türkiye
[2] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, 07059 Antalya, Türkiye
[3] Department of Animal Science, Faculty of Agriculture, Hatay Mustafa Kemal University, 31060 Hatay, Türkiye

Received 25 August 2024 – Revised 5 December 2024 – Accepted 10 December 2024

Abstract – This study aims to classify honey bee genotypes by integrating explainable artificial intelligence techniques, particularly decision trees, with both morphometric and molecular data. A total of 4949 samples were collected from 500 colonies across five regions in Türkiye, representing diverse subspecies and ecotypes. Morphometric data included 16 key wing characteristics, while molecular data contained 26 highly informative SSR loci. First, we used 16 morphometric wing parameters to classify bees into five regions where they originate. The decision tree algorithm resulted in a tree with wing length and O26 and L13 angles, but the classification accuracy was low (51%). Later, we included 26 molecular variables and obtained a decision tree with four SSR loci—Ap218, Ap274, Ap001, and Ap289—and achieved a high classification accuracy of 96.38%. The findings also revealed the first-ever identification of a SSR locus (Ap218) strongly associated with wing length in honey bees. Finally, we explained wing length with molecular data by modeling a regression decision tree. This tree identified Ap218, Ap223, and Ap001 as the most significant SSR loci for the wing length model. This study provides a powerful approach for differentiating honey bee genotypes while offering valuable insights into the genetic factors influencing wing morphology. The results have significant implications for the conservation and sustainable management of honey bee genetic resources, particularly in regions like Türkiye where genetic diversity is at risk.

Honey bee / Morphometry / SSR / Explainable AI / Decision tree

1. INTRODUCTION

Honey bees (*Apis mellifera*) are among the most ecologically significant pollinators, playing a vital role in sustaining biodiversity and supporting agricultural ecosystems worldwide (Gallai et al. 2009). Their role in the pollination of a wide range of crops and wild plants directly influences global food security and the maintenance of diverse ecosystems. The preservation of honey bee biodiversity, therefore, is of paramount importance, not just for ecological resilience but also for ensuring the long-term productivity and sustainability of agricultural systems. However, honey bees are increasingly under threat from a variety of factors, including habitat loss, climate change, pesticide exposure, and the introduction of non-native species (Grozinger and Zayed 2020). These threats have underscored the urgent need for effective conservation strategies to protect and sustain honey bee populations globally (Espregueira et al. 2020; Hristov et al. 2020; Panziera et al. 2022).

The precise identification and classification of honey bee genotypes is a critical component of conservation efforts. Accurate genotype

Corresponding author: U. Bilge, ubilge@akdeniz.edu.tr
Manuscript editor: Cedric Alaux.

Published online: 02 January 2025

classification not only aids in maintaining the genetic diversity essential for honey bee populations to adapt to changing environmental conditions but also helps to safeguard the evolutionary potential of locally adapted populations by preserving unique genetic traits and reducing the risk of genetic homogenization. Achieving this requires the implementation of reliable methods to differentiate between local and introduced genotypes, thereby providing crucial insights for breeding programs and guiding effective conservation policies. Traditional methods for identifying honey bee genotypes, such as mitochondrial DNA (mtDNA) analysis, restriction fragment length polymorphism (RFLP), and single nucleotide polymorphisms (SNP) (Kandemir and Kence 1995; Smith et al. 1997; Kekeçoğlu et al. 2009; Özdil et al. 2009; Chen et al. 2022), have been widely used due to their high accuracy. However, these methods are often resource-intensive and time-consuming (Henriques 2020). Another widely used method, simple sequence repeat (SSR) markers, is in principle included in this group of methods; however, when integrated with modern computational techniques, they have the potential to offer scalable, interpretable, and less resource-intensive solutions (Garcia et al. 2022; Yıldız et al. 2023).

In response to these challenges, more cost-effective and non-invasive alternatives have been developed, with morphometric analysis, particularly wing morphometry, becoming widely utilized in honey bee classification (De Nart et al. 2022). Morphometric methods, which involve the quantitative analysis of various anatomical features such as lengths, widths, distances between reference points, angles, and ratios, offer a practical solution. These methods, when analyzed using multivariate statistical techniques (Rohlf and Marcus 1993; Adams et al. 2004; Zelditch et al. 2012), can provide results that are often comparable to those obtained through molecular markers (Meixner et al. 2013; Oleksa et al. 2023). Wing morphometry, in particular, has proven to be highly informative, allowing for the differentiation of honey bee subspecies with a high degree of precision (Ruttner et al. 1978; Güler 2017).

The field of honey bee classification is becoming increasingly complex, with the integration of both morphological and molecular data. These data are inherently multivariate and involve intricate relationships that challenge the limits of traditional statistical methods (Crisci et al. 2012). This complexity has driven recent advancements in the field, leading to the adoption of artificial intelligence (AI) applications, particularly machine learning methods, to enhance the accuracy and efficiency of classification efforts. In a notable study, De Nart et al. (2022) utilized AI-based techniques to recognize honey bee wing images, aiming to differentiate between genotypes. Their analysis of 9887 wing images from 7 subspecies and 1 hybrid using advanced AI models, such as ResNet 50, MobileNet V2, Inception Net V3, and Inception ResNet V2, demonstrated that these models could achieve classification accuracy above 92%, outperforming traditional morphometric evaluations. Similarly, Rodrigues et al. (2022) developed the DeepWings© software, which leverages AI to automatically detect 19 landmarks on digital wing images and perform geometric morphometric classification. This software's implementation showed an average accuracy of 86.6% across 26 subspecies and 95.8% for a subset of five major subspecies, with a high processing speed that underscores its practical application in research settings.

Despite the high accuracy rates of AI-driven image processing methods, which require extensive datasets and powerful computational resources, a critical gap remains in the literature: the lack of AI studies focusing on the explainability of honey bee classification. Explainability in AI, particularly in the context of complex biological data, is crucial as it enables researchers to understand the underlying factors driving classification decisions, thus facilitating the integration of AI findings into broader scientific knowledge. Garcia et al. (2022) explored the relationship between morphometric predictions and molecular data in one of the few studies to do so, yet the potential of AI to provide explainable insights into honey bee subspecies classification has not been fully realized.

This study seeks to bridge this gap by employing explainable AI (XAI) techniques, such as decision trees, to classify honey bee genotypes using a combination of morphometric and molecular data. By analyzing 5000 samples from 500 colonies across five regions in Türkiye, representing various subspecies and ecotypes, we aim to enhance the accuracy of classification while also providing valuable insights into the most informative features for honey bee classification. Our approach not only aims to improve the understanding of honey bee genotype differentiation but also has significant implications for the conservation and sustainable management of honey bee genetic resources. The integration of XAI techniques offers the potential to refine data collection and analysis processes, ultimately contributing to more effective and efficient conservation strategies.

2. MATERIALS AND METHODS

2.1. Sample collection

The molecular and morphometric data of honey bees were obtained from the project coded "TAGEM-18 AR-GE 07" which was supported by the Republic of Türkiye Ministry of Agriculture and Forestry in 2018. Within the scope of this project, bees representing different honey bee subspecies and ecotypes were collected from beekeeping centers established in the following regions: Hatay (*A. m. syriaca*), Kırşehir (*A. m. anatoliaca*), Düzce (*A. m. anatoliaca*, Yığılca ecotype), Kırklareli (*A. m. carnica*), and Çanakkale (*A. m. anatoliaca*, Gökçeada ecotype).

Care was taken to select samples of young worker bees that had not participated in external activities, representing each colony. The samples were collected during the swarming period to ensure that the individual honey bees were in optimal morphological condition, characterized by well-developed body structures such as intact wings and a healthy, undamaged exoskeleton. Young worker bees were gently brushed into plastic bags from the sealed brood area of the colonies using a soft brush. A total of 5000 individual bees were collected, with 10 individuals from each of the 500 colonies, comprising 100 colonies from each province ($5 \times 100 \times 10 = 5000$). The bees were numbered and stored in centrifuge tubes containing 70% ethanol. Upon arrival at the laboratory, the samples were air-dried at room temperature for 1 h to allow the ethanol to evaporate (Fig. 1).

2.2. Molecular data

Individual DNA extractions from worker bees were performed using the High Pure PCR Template Preparation Kit (Roche, Basel, Switzerland). Genomic variations among DNA samples were estimated using 26 highly informative SSR loci, as previously described in the literature (Solignac et al. 2003). The thermal cycling for PCR was carried out as follows: initial denaturation at 94 °C for 5 min, followed by 35 cycles of 94 °C for 30 s, 54–61 °C (specific annealing temperature of the primer) for 30 s, and 72 °C for 45 s, with a final extension step at 72 °C for 5 min. The lengths of the DNA fragments corresponding to the SSR loci were precisely determined using an AATI fragment analyzer. The obtained fragment lengths were analyzed using the PROSize v.2.0 software. Pairwise genetic differentiation between populations was assessed using the Arlequin software package (Excoffier et al. 2007). Genetic diversity parameters, including N (number of loci), N_P (number of polymorphic loci), N_A (observed number of alleles), N_E (effective number of alleles), ASR (allelic size range), H_E (expected heterozygosity), H_O (observed heterozygosity), G-W (Garza-Williamson Index), Theta (H, molecular diversity indices), and F_{IS} (inbreeding coefficient within groups), were calculated using Popgene v.1.32 (Yeh et al. 1997) and Arlequin v.3.11 (Excoffier et al. 2007). PIC values, which provide information on the usefulness of markers, were calculated using the Microsatellite Toolkit (Park 2008).

2.3. Morphometric data

The biometric measurement of morphological characteristics was conducted according to Ruttner et al. (1978) and Güler (2017). During the measurements, an ocular micrometer

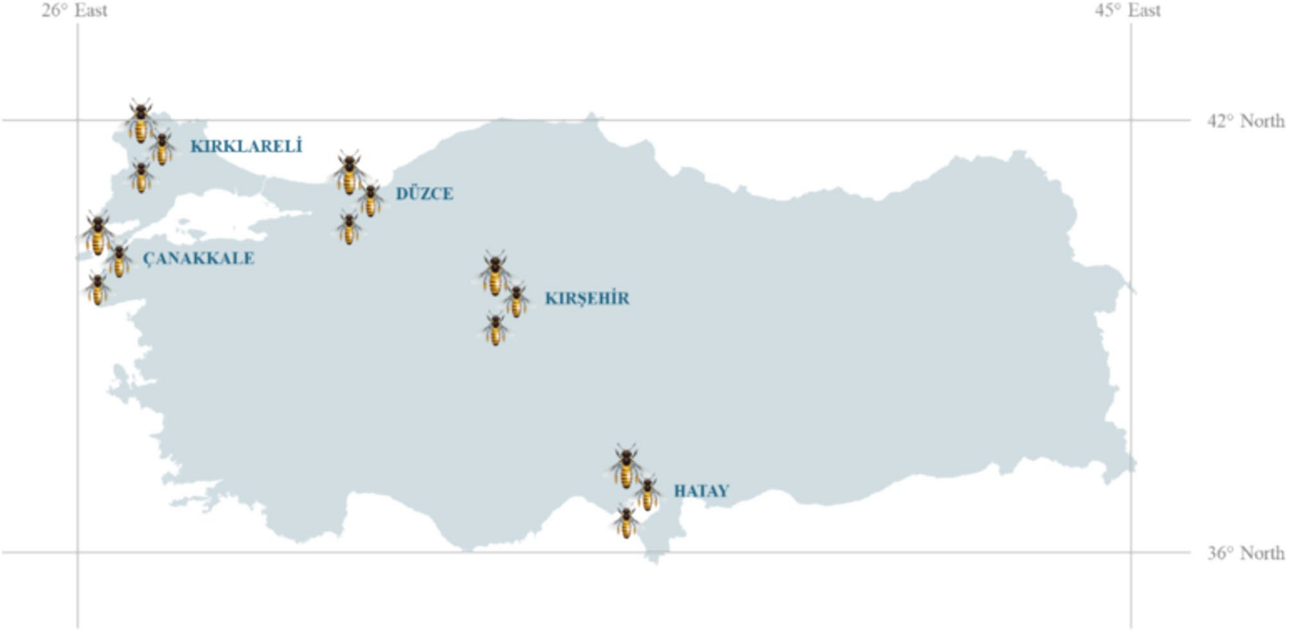


Figure 1. Provinces where honey bee samples were collected.

was attached to the binocular microscope, and a magnification factor of × 2 was applied. The measurements obtained were converted to the metric system (mm) by multiplying them by a coefficient of 0.0574 for × 2 magnification and 0.0271 for × 4 magnification.

The right wings of the samples were carefully detached from the thorax at the junction with the forefinger and thumb of the right hand and then mounted on slides coated with Hoyer's solution. Figure 2 illustrates the wing length and wing width (Güven 2003).

The vein length between the 3rd cubital cell and the discoidal cell on the right forewing of worker bees was measured as "a," and the vein length "b," which forms an angle of 151° with this vein, was also measured. The Cubital Index was then calculated by determining the ratio of the "a" vein length to the "b" vein length (a/b). Additionally, 11 wing vein angles located on the right forewing, specifically 1-A4, 2-B4, 3-D7, 4-E9, 5-G18, 6-J10, 7-J16, 8-K19, 9-L13, 10-N23, and 11-O26 (Fig. 3), were measured using a stereomicroscope equipped with a drawing tube attachment. The drawing tube was used to mark the points of the wing vein angles on an A4 paper sheet. These points were then connected to form the sides of the angles, and the angle values were determined using a protractor (Güler and Toy 2008). Figure 3 illustrates the vein angles measured on the forewing of a worker bee and their positions on the wing.

2.4. Explainable AI

There were several machine learning algorithms at our disposal such as random forest and gradient boost algorithms. As our focus is not to predict which subpopulation a given bee belongs to, but to identify the dynamics of morphological and molecular differences between bees in different subpopulations, we chose the decision tree technique as it has a high explanation capability (Islam et al. 2021).

2.4.1. Data processing and analysis

Missing data for each population were imputed using the mean value of the same population. The optimization and hyperparameter tuning of the decision tree model were carried out using appropriate methods, such as cross-validation or grid search.

2.4.2. Decision trees

Decision trees were employed for classification and analysis based on the features in the dataset. These analyses were conducted using version 4.3.0 of the R programming language and the classical "rpart" package (R Core Team 2024). The splitting points were automatically

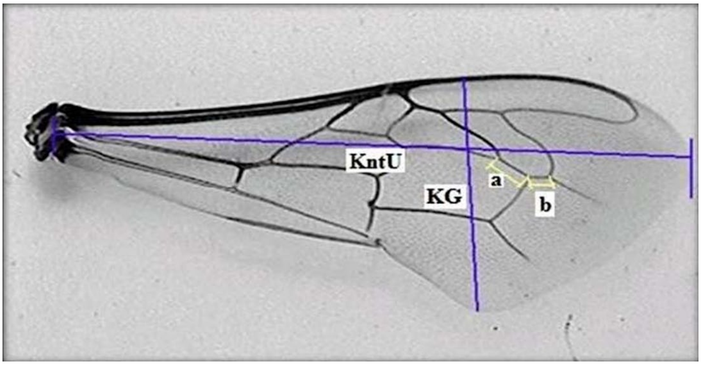


Figure 2. Wing dimensions on the forewing of a worker bee are KntU: wing length, KG: wing width, and a: cubital a and b: cubital b wing vein length (Güven 2003).

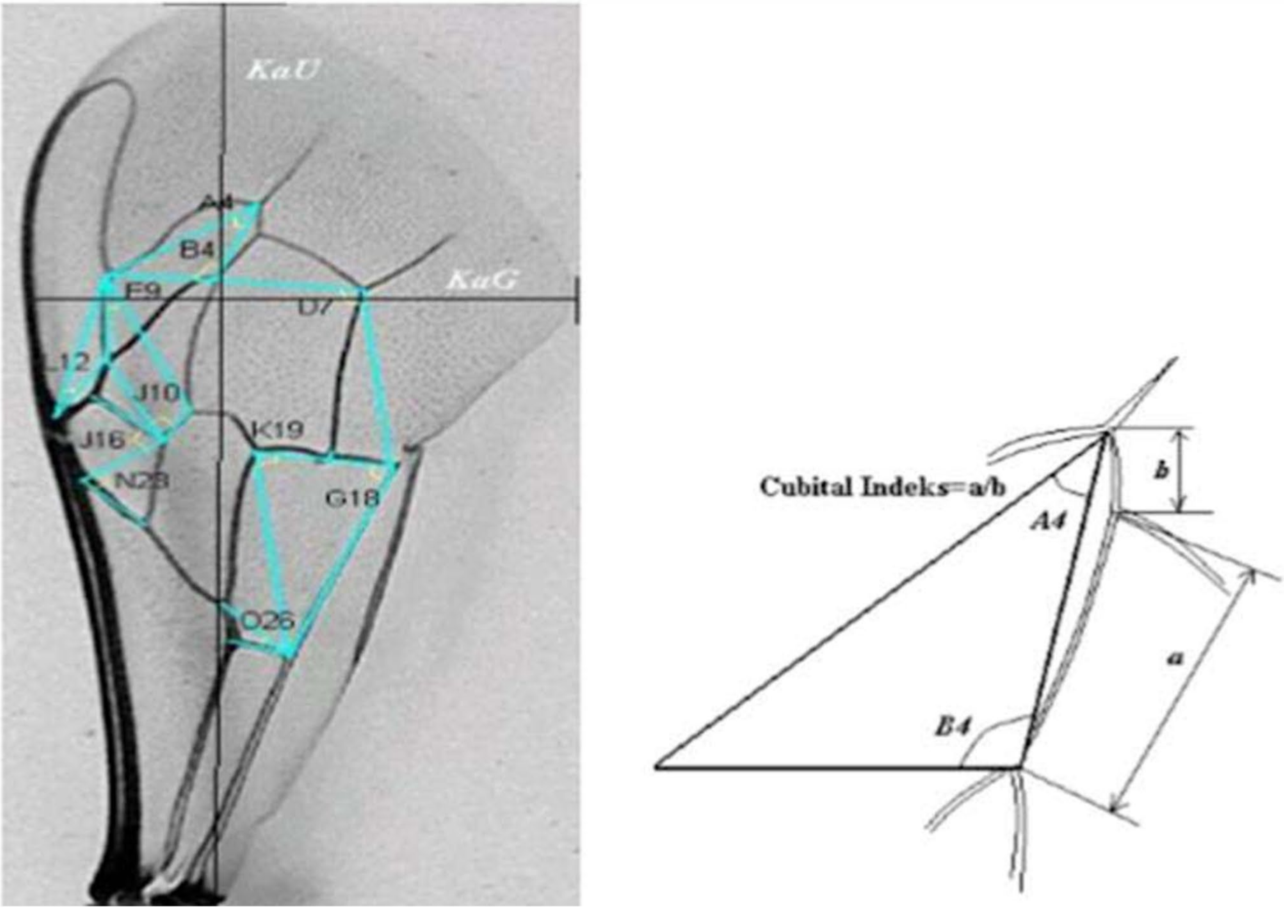


Figure 3. On the left, the forewing and forewing vein angles of the worker bee are shown: 1-A4, 2-B4, 3-D7, 4-E9, 5-G12, 6-J10, 7-J16, 8-K19, 9-L13, 10-N23, and 11-O26. On the right, the measurement of the Cubital Index character is shown: a = cubital vein a and b = cubital vein b (Güler 2017).

determined through the optimization features provided by the package. For classification based on morphometric data, decision trees were constructed with the parameters minbucket = 100 and maxdepth = 3. Similarly, the classification analysis using SSR data was conducted with the same parameters. To explain wing length using molecular data, decision trees were constructed with the parameters method = "anova," minbucket = 50, and maxdepth = 4. In these analyses, the entire dataset was used to prioritize explanatory power over predictive performance.

2.4.3. *Confusion matrix*

A confusion matrix was employed to evaluate the classification performance. This matrix illustrates the relationship between the predicted classes and the actual classes. In R, the confusion matrix was calculated using appropriate packages, such as "caret" or "e1071," and performance metrics such as accuracy, precision, and specificity were analyzed.

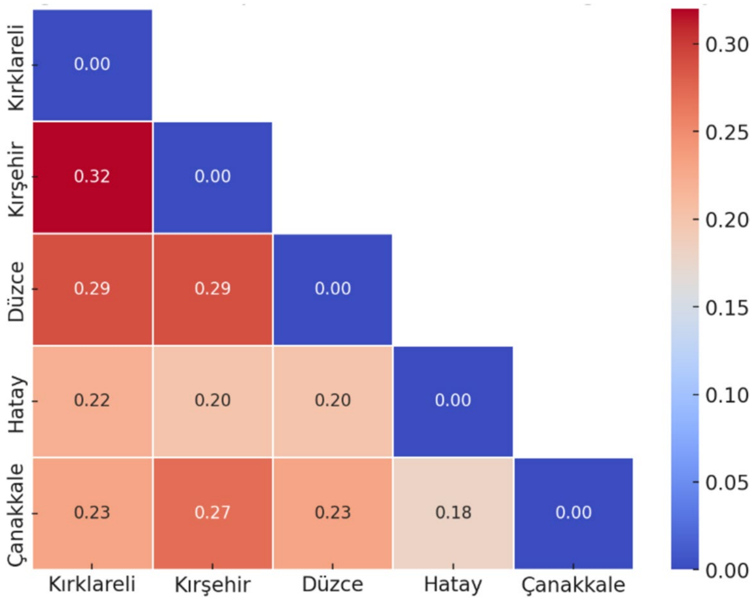
3. RESULTS AND DISCUSSION

In the present study, a total of 5000 individuals from five different provinces in Türkiye (Hatay (*A. m. syriaca*), Kırşehir (*A. m. anatoliaca*), Düzce (*A. m. anatoliaca*, Yığılca ecotype), Kırklareli (*A. m. carnica*), and Çanakkale (*A. m. anatoliaca*, Gökçeada ecotype)) were analyzed. However, some samples were excluded from the study due to concerns regarding their reliability, such as physical damage to the specimens, incomplete data, or inconsistencies in morphometric measurements. As a result, the study continued with 4949 samples. Pairwise F_{ST} values for the honey bee populations examined ranged from 0.18 to 0.32 (Fig. 4). The lowest genetic differentiation coefficient was observed between the Hatay and Çanakkale populations (0.18), while the highest F_{ST} value was identified between the Kırklareli and Kırşehir populations (0.32). Overall, moderate genetic variation was observed among the studied populations. All pairwise F_{ST} values found in this study were determined to be statistically significant ($p < 0.05$).

In this study, genetic variation parameters and inbreeding coefficient (F_{IS}) values were calculated based on 26 SSR loci, and the results are summarized in Table I. Upon examining the table, it is evident that all 26 SSR loci studied for the five populations are polymorphic. The PIC values being greater than 0.5 indicate that the selected SSR loci have high information content for detecting polymorphism. A PIC value greater than 0.5 demonstrates that the marker provides a high level of information, while a PIC value above 0.75 indicates that the locus is highly informative and much more useful for genetic variation and genetic mapping studies (Boltstein et al. 1980). This finding confirms that the loci selection was appropriate for revealing the genetic diversity in the studied populations (Table I). The lowest N_A, N_E, PIC, H_E, G-W Index, and Theta (H) values were observed in the Kırklareli population. Observed heterozygosity in the Kırşehir population was found to be higher compared to the other populations. The highest within-group variation (F_{IS}) was identified in the Hatay honey bee population, while the lowest was found in the Kırşehir population.

In the one-way analysis of variance, it was found that honey bees from the five regions significantly differed from each other in terms of 16 morphological characteristics ($p < 0.001$). When these bees were morphologically characterized in sequence, the Çanakkale bees had the smallest E9 angle and the largest wing length (WL), as well as the largest A4, D7, G12, J10, K16, and N23 angles; the Düzce bees had the largest wing width (WW) and Cubital Index (CI); the Hatay bees had the smallest wing length (WL), wing width (WW) and K16 and O26 angles; the Kırklareli bees had the smallest A4 and D7 angles and the largest B4, E9, and J16 angles; and the Kırşehir bees had the smallest Cubital Index (CI) and B4, G12, J10, J16, and N23 angles and the largest L13 and O26 angles (Table II).

Using the decision tree algorithm for classification based on morphological data, the characters WL, L13, and O26 were found to be the most



	Kırklareli	Kırşehir	Düzce	Hatay	Çanakkale
Kırklareli	0.00				
Kırşehir	0.32	0.00			
Düzce	0.29	0.29	0.00		
Hatay	0.22	0.20	0.20	0.00	
Çanakkale	0.23	0.27	0.23	0.18	0.00

Figure 4. Pairwise F_{ST} values of populations (distance method: pairwise differences).

Table I Genetic diversity parameters based on SSR markers

Parameters	Provinces					Means
	Kırklareli	Kırşehir	Düzce	Hatay	Çanakkale	
N	26	26	26	26	26	26
N_P	26	26	26	26	26	26
N_A	11.76 ± 3.84	19.89 ± 6.07	14.06 ± 5.52	17.43 ± 5.23	12.72 ± 4.18	15.17 ± 4.96
N_E	5.54 ± 3.36	9.39 ± 4.60	6.05 ± 2.94	9.78 ± 3.91	7.35 ± 3.12	7.62 ± 3.58
ASR	10.53 ± 7.46	16.48 ± 7.17	11.10 ± 6.31	13.83 ± 5.29	10.24 ± 5.20	12.43 ± 6.29
PIC	0.743	0.830	0.776	0.865	0.790	0.801
H_E	0.77 ± 0.10	0.87 ± 0.04	0.80 ± 0.09	0.88 ± 0.06	0.84 ± 0.09	0.83 ± 0.07
H_O	0.16 ± 0.13	0.22 ± 0.17	0.14 ± 0.16	0.12 ± 0.15	0.16 ± 0.18	0.16 ± 0.15
G-W	0.76 ± 0.22	0.78 ± 0.18	0.80 ± 0.16	0.79 ± 0.16	0.79 ± 0.20	0.78 ± 0.18
Theta (*H*)	1.68	2.31	1.86	2.84	1.96	2.13
F_{IS}	0.78	0.75	0.82	0.89	0.80	0.80

N number of loci, N_P number of polymorphic loci, N_A observed number of alleles, N_E effective number of alleles, *ASR* allelic size range, *PIC* polymorphic information content, H_E expected heterozygosity, H_O observed heterozygosity, *G-W* Garza-Williamson Index, *Theta (H)* molecular diversity indices, F_{IS} inbreeding coefficient within groups.

distinguishing features among the genotypes out of the total 16 characteristics. In the constructed decision trees, each node represents a feature and a threshold value that divides the data into two branches. A "Yes" answer to the threshold condition leads to the left branch, while a "No" answer directs to the right branch. For example, individuals with a wing length of 8.7 mm or

Table II Mean and standard error values of morphological characters determined by direct measurement of bee samples representing Çanakkale, Düzce, Hatay, Kırklareli, and Kırşehir bee genotypes

Characters	Provinces					Means
	Hatay	Kırklareli	Kırşehir	Çanakkale	Düzce	
WL	8.295 ± 0.132	8.317 ± 0.096	8.376 ± 0.090	9.072 ± 0.213	9.056 ± 0.181	8.621 ± 0.391
WW	2.801 ± 0.061	2.832 ± 0.056	2.821 ± 0.048	3.102 ± 0.082	3.135 ± 0.078	2.937 ± 0.162
A	0.487 ± 0.022	0.493 ± 0.023	0.48 ± 0.02	0.514 ± 0.031	0.539 ± 0.028	0.502 ± 0.033
B	0.239 ± 0.014	0.241 ± 0.012	0.25 ± 0.017	0.246 ± 0.013	0.242 ± 0.015	0.244 ± 0.015
CI	2.07 ± 0.172	2.083 ± 0.143	1.958 ± 0.163	2.115 ± 0.162	2.267 ± 0.217	2.098 ± 0.199
A4	31.409 ± 1.386	31.303 ± 1.169	32.936 ± 1.335	32.987 ± 1.257	32.53 ± 1.039	32.225 ± 1.441
B4	103.835 ± 3.426	105.167 ± 3.328	102.152 ± 2.736	103.267 ± 2.918	102.458 ± 2.983	103.388 ± 3.263
D7	100.951 ± 1.773	100.605 ± 1.78	101.655 ± 2.14	102.168 ± 2.05	100.881 ± 1.737	101.249 ± 1.98
E9	20.108 ± 0.939	20.95 ± 1.080	20.416 ± 1.105	19.236 ± 0.639	19.342 ± 0.739	20.015 ± 1.123
G12	88.915 ± 1.883	88.718 ± 2.079	88.477 ± 2.211	92.991 ± 2.322	91.969 ± 1.722	90.205 ± 2.781
J10	49.239 ± 1.778	49.888 ± 1.687	48.5 ± 2.038	52.269 ± 1.914	51.927 ± 2.016	50.361 ± 2.396
J16	93.266 ± 2.479	94.497 ± 2.466	92.128 ± 2.198	93.741 ± 2.793	92.297 ± 2.275	93.195 ± 2.599
K16	74.378 ± 1.803	74.747 ± 1.717	75.68 ± 1.936	77.202 ± 2.041	76.352 ± 1.757	75.664 ± 2.117
L13	15.8 ± 0.837	16.563 ± 1.099	17.775 ± 0.97	15.403 ± 0.899	15.304 ± 0.988	16.168 ± 1.325
N23	90.147 ± 2.27	91.808 ± 2.952	88.52 ± 2.789	94.235 ± 2.625	93.581 ± 2.266	91.658 ± 3.341
O26	32.874 ± 1.676	34.600 ± 1.964	74.165 ± 377.444	38.059 ± 2.675	35.871 ± 2.325	43.011 ± 167.819

greater are directed to the left branch, whereas those with a wing length below this threshold are directed to the right branch. This hierarchical structure systematically divides the population into smaller, more homogeneous groups, as seen in Fig. 5. The terminal nodes ("leaves") summarize the classification outcome, such as the number of individuals within a particular group or their predominant origin. According to this, individuals with a wing length of 8.7 mm or greater comprised 39% of the entire population, with this group being primarily composed of bees from Çanakkale (949) and Düzce (945). Individuals with an O26 angle of 38 or greater were classified into a smaller group, representing 15% of the entire population, and within this group, 461 were from Çanakkale (Fig. 5). While the Cubital Index, fore wing length, and wing venation angles have traditionally been considered the most common and reliable parameters for studying honey bee biodiversity (Ruttner et al. 1978; Güler 2017), our findings indicate that fore wing length (WL), in combination with other morphological features such as O26 and L23, may also be valuable markers for distinguishing honey bee genotypes. In particular, the fore wing length is a significant characteristic in honey bee biology, as it is often associated with key aspects such as body size, developmental stage, and overall fitness (Shingleton et al. 2005; Dadgostar et al. 2020). These factors, in turn, can be influenced by both genetic and environmental factors, including nutrition and climate (Shingleton et al. 2005).

The classification accuracy significantly increased when molecular data was used with the decision tree algorithm, compared to when only morphological data was used. The loci Ap218, Ap274, Ap001, and Ap289 demonstrated a high level of distinctiveness among the genotypes. Accordingly, individuals with Ap218 locus length of 116 bp or more represented 60% of the entire population, with this group primarily composed of bees from Çanakkale (989), Düzce (970), and Kırklareli (1000). Within this group, individuals with an Ap274 locus length of 124 bp or more were classified into a smaller group representing 19% of the entire population, of which

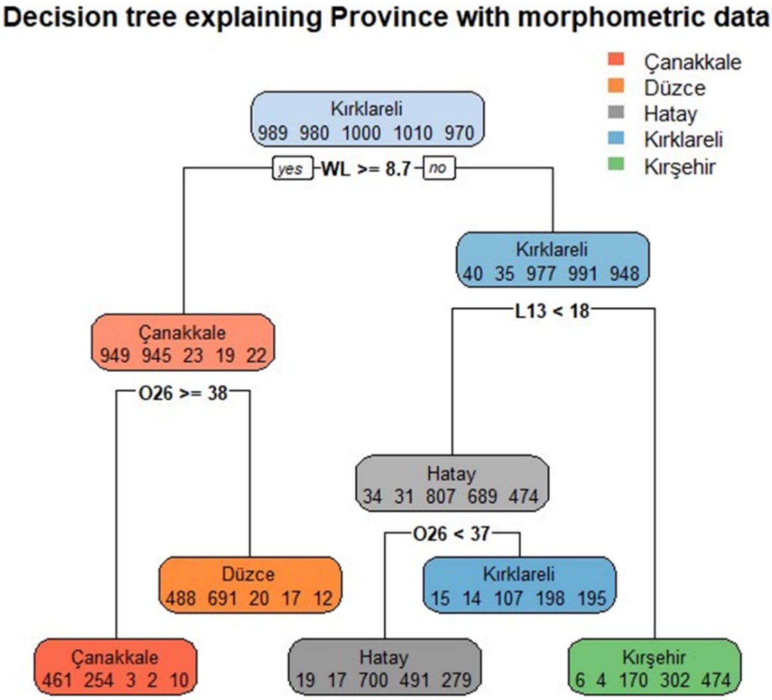


Figure 5. Morphometric classification of honey bees using decision tree algorithm. WL represents wing length, while L13 and O26 denote wing angles.

920 were from Çanakkale (Fig. 6). While the classification of bees is primarily based on morphological characters and morphometric measurements, it has often been suggested that these methods have the potential to produce misleading results (Gonçalves et al. 2022). Research has shown that molecular techniques have the capacity to differentiate almost all genotypes (Whitfield et al. 2006; Spötter et al. 2012; Momeni et al. 2021). Moreover, the decision tree algorithm greatly optimized the analysis process by selecting only 4 loci out of the 26 in the dataset. The high distinctiveness of the Ap218, Ap274, Ap001, and Ap289 loci in genotype classification suggests that results can be obtained quickly and cost-effectively without the need to analyze most other loci. This approach not only saves time but also reduces costs, thereby enhancing research efficiency.

The classification predictions using morphometric and molecular data of the genotypes, along with the confusion matrix, are presented in Fig. 7. The rows represent the reference data, while the columns represent the model predictions. The breeds were classified with high accuracy (0.9661 accuracy within a 95% confidence interval), and the matrix was obtained as diagonal. The low level of off-diagonal cells in the confusion matrix indicates that there were very few errors in the predictions. The study included populations with differences not only at the subspecies level but also at the ecotype level (Düzce and Çanakkale), making this accuracy particularly important. This is significant because ecotypes are expected to be more challenging to differentiate due to their lower degree of differentiation compared to subspecies. In the literature, AI-based classifications typically involve populations with differences at the subspecies level (Bustamante et al. 2021; De Nart et al. 2022; Rodrigues et al. 2022; Garcia et al. 2022). Previous studies have

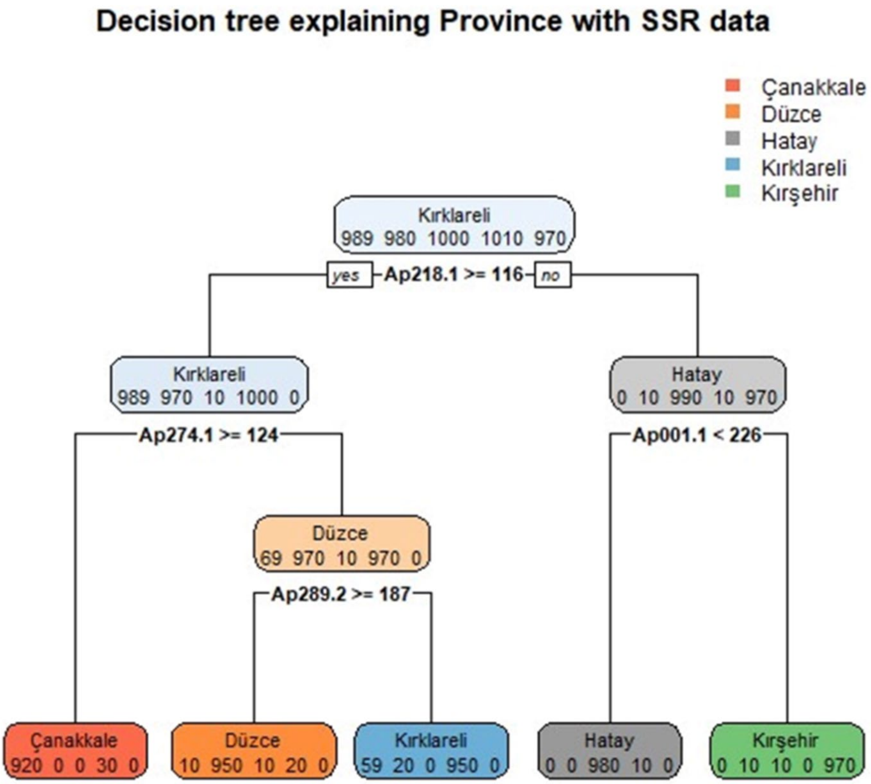
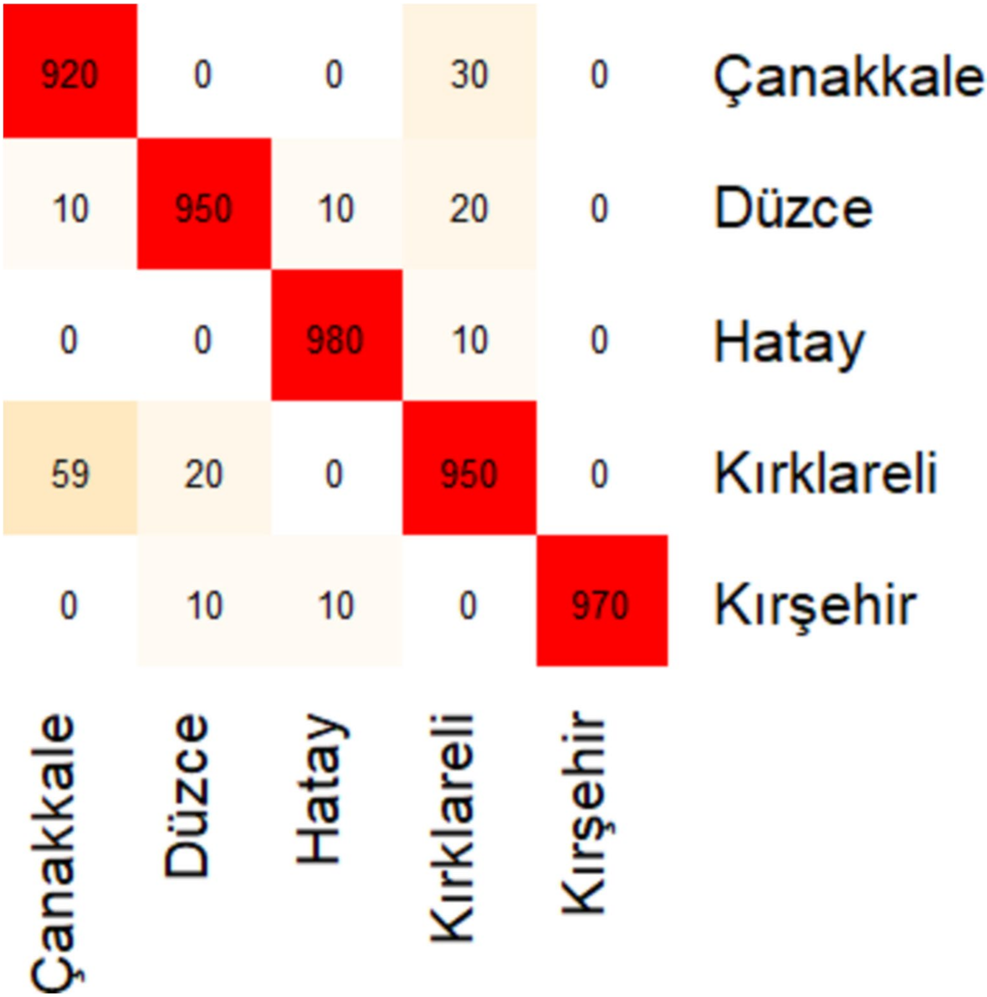


Figure 6. Molecular classification of honey bees using decision tree algorithm. Ap218, Ap274, Ap001, and Ap289 denote SSR locus names.

frequently mentioned the difficulties in classifying subspecies in Türkiye. For example, factors such as migratory beekeeping practices, flawed breeding programs, and uncontrolled distribution of commercial queens in Türkiye have led to changes in the genetic diversity of local honey bee genotypes, making classification more challenging (Güler and Toy 2008; Karabağ et al. 2020; Yıldız et al. 2023).

To understand the genetic foundations of honey bee wing traits and to examine the effects of SSR loci on phenotypic variability, the loci influencing the WL trait were investigated using the decision tree algorithm. The SSR loci identified as the most important determinants in the decision tree results were Ap218, Ap223, and Ap001. These loci appeared as the top-level branching points in the model and played a critical role in predicting wing length. In particular, the Ap218 locus provided the highest information gain compared to the other markers and was found to be associated with wing length in honey bees. Our results indicate that the presence of the Ap218 locus with a length greater than 120 base pairs was correlated with wing lengths exceeding 9 mm (Fig. 8). While this is the first time an SSR locus associated with wing length has been identified in honey bees, it is important to note that this correlation does not necessarily imply a direct causal relationship. The Ap218 locus may serve as a breed-specific marker, reflecting genetic variation that could influence morphological traits like wing length, but it might not directly determine the trait itself. Additionally, the observed association may be influenced by parallel processes, such as



	Çanakkale	Düzce	Hatay	Kırklareli	Kırşehir
Çanakkale	920	0	0	30	0
Düzce	10	950	10	20	0
Hatay	0	0	980	10	0
Kırklareli	59	20	0	950	0
Kırşehir	0	10	10	0	970

Figure 7. Classification predictions using confusion matrix with morphometric and molecular data. Rows show reference data and columns show predicted data.

environmental factors or other genetic interactions, which could also affect wing morphology. No previous studies in the literature have reported genetic markers directly linked to honey bee wing size, making our finding an important contribution to understanding the genetic structure of honey bees. This discovery opens new avenues for future morphological and genetic studies, though further research is needed to confirm the precise role of the Ap218 locus and explore other potential contributing factors.

It would not be incorrect to say that the predictions obtained using next-generation machine learning algorithms based on image processing for the classification of honey bee genotypes generally exhibit high accuracy (Bustamente et al. 2021; De Nart et al. 2022; Rodrigues et al. 2022; Garcia et al. 2022). This is because computerized image recognition systems have the ability to deliver better results than the human eye, even with blurred or damaged wing images (De Nart et al. 2022). The study with the highest

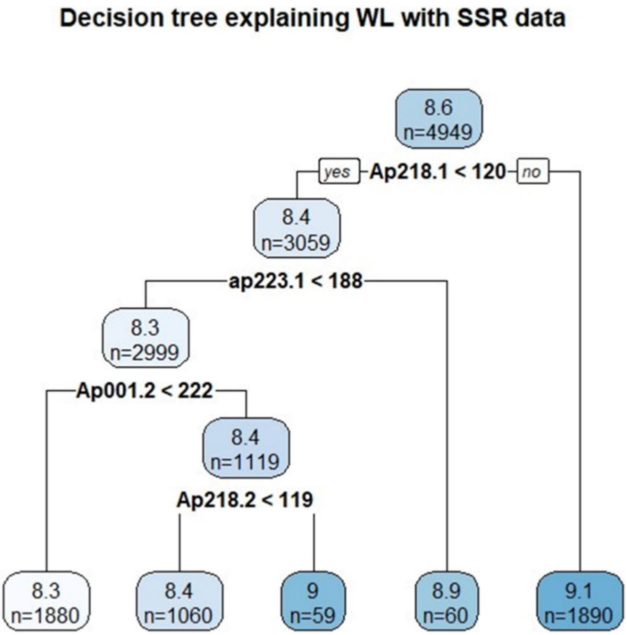


Figure 8. Determination of important SSR loci affecting wing length using decision tree model. Ap218, Ap223, Ap001, and Ap274 denote SSR locus names.

prediction accuracy among current research was conducted by Bustamante et al. (2021). In their study, they developed a software using geometric morphometry techniques to classify honey bees at the species level. They photographed a total of 8756 wings from individuals belonging to 9 Apis species. As a result, they developed a web application capable of correctly identifying 99.4% of individuals based on a single wing image at the species level (http://wing-app-client.s3-website.us-east-2.amazonaws.com).

However, it has also been reported that more advanced and complex machine learning methods, such as image processing, are not always more accurate than traditional machine learning methods (Blair et al. 2022). In their study on classifying invertebrates, Blair et al. (2022) reported that the XGBoost algorithm produced results quite similar to those of the ResNet-50 CNN, despite having less data. Moreover, the complexity and decision-making processes of image processing-based models are often opaque, making it difficult to clearly understand which features are effective in classification. This results in decision-making processes often being viewed as a "black box" contrasting with the more understandable traditional algorithms (Islam et al. 2021). Due to neurons in deep layers that identify highly abstract features, it is challenging to determine which features and patterns contribute to the classification process. This makes it difficult to evaluate the model's reliability and accuracy because identifying and correcting error sources in cases where the model makes incorrect decisions are complex processes. Additionally, the lack of explainability in deep learning models presents a significant disadvantage in fields such as biological and ecological research. In studies like honey bee wing classification, researchers not only aim for accurate classification but also seek to understand the biological meaning and factors underlying the classification. While traditional methods offer greater explainability in this regard, the limitations of CNNs in providing this information complicate the in-depth analysis of scientific research and the interpretation of results.

In our study, we emphasize the use of the decision tree algorithm as one of the easiest ways to achieve explainable results for honey bee classification. By using decision trees, we have, for the first time, clearly determined the importance levels of both morphometric and molecular parameters used in honey bee classification. The explainability provided by decision trees enhances the model's reliability and facilitates the in-depth analysis of scientific research.

4. CONCLUSION

In conclusion, the integration of explainable artificial intelligence (XAI) techniques, particularly decision trees, with morphometric and molecular data offers a robust approach for the classification of honey bee genotypes. This study not only demonstrates high classification accuracy but also provides valuable insights into the key genetic factors influencing wing morphology. These findings have significant implications for the conservation and sustainable management of honey bee genetic resources, especially in regions like Türkiye where genetic diversity is under threat.

The approach developed here serves as a potential model for future studies aiming to balance accuracy with interpretability in biological classification tasks. The ability to clearly understand and explain the factors driving classification decisions is crucial, particularly in ecological and biological research where the underlying biological significance of the results is just as important as the accuracy of the predictions.

Furthermore, this study has, for the first time, identified specific SSR loci that are strongly associated with wing length in honey bees, adding a new dimension to our understanding of the genetic architecture underlying key phenotypic traits. The use of decision trees has highlighted the potential of these loci to streamline and enhance the efficiency of genetic studies, offering a cost-effective and time-saving alternative to more complex models.

In conclusion, the findings of this research underscore the importance of incorporating explainable models in the study of honey bee genetics and morphology, paving the way for more informed and effective conservation strategies. By offering a clearer understanding of the genetic and phenotypic diversity within honey bee populations, this study contributes to the broader goal of preserving biodiversity and ensuring the sustainability of vital pollinator species.

AUTHOR CONTRIBUTION

BİY and KK designed and planned the study and drafted the manuscript. AG was responsible for data collection. UB handled the data analysis. The final manuscript was reviewed and approved by all authors.

FUNDING

No funding was received.

DATA AVAILABILITY

The authors confirm that the data and materials supporting the results of the study are available within the article.

CODE AVAILABILITY

Not applicable.

DECLARATIONS

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

REFERENCES

Adams DC, Rohlf FJ, Slice DE (2004) Geometric morphometrics: ten years of progress following the 'revolution'. Ital J Zool 71(1):5–16

Blair J, Weiser MD, de Beurs K, Kaspari M, Siler C, Marshall KE (2022) Embracing imperfection: machine-assisted invertebrate classification in real-world datasets. Ecol Informatics 72:101896

Bolstein D, White RL, Skolnik M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32(3):314–331

Bustamante T, Fuchs S, Grünewald B et al (2021) A geometric morphometric method and web application for identifying honey bee species (Apis spp.) using only forewings. Apidologie 52:697–706

Chen C, Parejo M, Momeni J et al (2022) Population structure and diversity in European honey bees (*Apis mellifera* L.) - an empiricalcomparison of pool and individual whole-genome sequencing. Genes 13(2):182

Crisci C, Ghattas B, Perera G (2012) A review of supervised machine learning algorithms and their applications to ecological data. Ecol Model 240:113–122

Dadgostar S, Nozari J, Tahmasbi G (2020) Wing characters for morphological study on the honey bee (*Apis mellifera* L.) populations among six provinces of Iran. Arthropods 9(4):129–138

De Nart D, Costa C, Di Prisco G, Carpana E (2022) Image recognition using convolutional neural networks for classification of honey bee subspecies. Apidologie 53(1):5

Espregueira Themudo G, Rey-Iglesia A, Robles Tascón L, Bruun Jensen A, da Fonseca RR, Campos PF (2020) Declining genetic diversity of European honeybees along the twentieth century. Sci Rep 10(1):10520

Excoffier L, Laval G, Schneider S (2007) ARLEQUIN (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 23(1):47–50

Gallai N, Salles JM, Settele J, Vaissière BE (2009) Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecol Econ 68(3):810–821

Garcia CAY, Rodrigues PJ, Tofilski A, Elen D, McCormak GP, Oleksa A et al (2022) Using the software DeepWings© to classify honey bees across Europe through wing geometric morphometrics. Insects 13(12):1132

Gonçalves LT, Françoso E, Deprá M (2022) Shorter, better, faster, stronger? Comparing the identification performance of full-length and mini-DNA barcodes for apid bees (Hymenoptera: Apidae). Apidologie 53(5):55

Grozinger CM, Zayed A (2020) Improving bee health through genomics. Nat Rev Genet 21(5):277–291

Güler A, Toy H (2008) Sinop ili Türkeli yöresi bal arıları (*Apis mellifera* L.)'nın morfolojik özellikleri. Anadolu J Agric Sci 23(3):190–197

Güler A (2017) Bal arısı (*Apis mellifera* L.) yetiştiriciliği hastalıkları ve ürünleri. Bereket Academy Publications, Ankara

Güven H (2003) Kuzeydoğu Anadolu ve Karadeniz Bölgesi'ndeki bazı arı (*Apis mellifera* L.) genotiplerinin morfolojik özellikleri ve performanslarının belirlenmesi. Dissertation, Samsun Ondokuz Mayıs University

Henriques D, Chávez-Galarza J, Teixeira JSG, Ferreira H, Neves CJ, Francoy TM, Pinto MA (2020) Wing geometric morphometrics of workers and drones and single nucleotide polymorphisms provide similar genetic structure in the Iberian honey bee (*Apis mellifera iberiensis*). Insects 11(2):89

Hristov P, Shumkova R, Palova N, Neov B (2020) Factors associated with honey bee colony losses: a mini-review. Vet Sci 7(4):166

Islam SR, Eberle W, Ghafoor SK, Ahmed M (2021) Explainable artificial intelligence approaches: a survey. arXiv. https://doi.org/10.48550/arXiv.2101.09429

Kandemir I, Kence A (1995) Allozyme variability in a central Anatolian honey bee (*Apis mellifera* L.) population. Apidologie 26(6):503–510

Karabağ K, Tunca Rİ, Tüten E, Doğaroğlu T (2020) Current genetic status of honey bees in Anatolia in terms of thirty polymorphic microsatellite markers. Turk J Entomol 44(3):333–346

Kekeçoğlu M, Bouga M, Harizanis P, Soysal MI (2009) Genetic divergence and phylogenetic relationships of honey bee populations from Türkiye using PCR-RFLP's analysis of two mtDNA segments. Bulgar J Agric Sci 15(6):589–597

Meixner MD, Pinto MA, Bouga M, Kryger P, Ivanova E, Fuchs S (2013) Standard methods for characterising subspecies and ecotypes of *Apismellifera*. J Apic Res 52(4):1–28

Momeni J, Parejo M, Nielsen RO, Langa J, Montes I, Papoutsis L et al (2021) Authoritative subspecies diagnosis tool for European honey bees based on ancestry informative SNPs. BMC Genomics. https://doi.org/10.1186/s12864-021-07379-7

Oleksa A, Căuia E, Siceanu A, Puškadija Z, Kovačić M, Pinto MA et al (2023) Honey bee (*Apis mellifera*) wing images: a tool for identification and conservation. GigaScience 12:giad019

Özdil F, Yildiz MA, Hall HG (2009) Molecular characterization of Turkish honey bee populations (*Apis mellifera* L.) inferred from mitochondrial DNA RFLP and sequence results. Apidologie 40(5):570–576

Panziera D, Requier F, Chantawannakul P, Pirk CW, Blacquière T (2022) The diversity decline in wild and managed honey bee populations urges for an integrated conservation approach. Front Ecol Evol 10:767950

Park SDE (2008) Excel microsatellite toolkit. Computer program and documentation distributed by the author. http://animalgenomics.ucd.ie/sdepark/ms-toolkit/. Accessed 20 Mar 2012

R Core Team (2024) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 18 Dec 2024

Rodrigues PJ, Gomes W, Pinto MA (2022) DeepWings©: automatic wing geometric morphometrics classification of honey bee (*Apis mellifera*) subspecies using deep learning for detecting landmarks. Big Data Cogn Comput 6(3):70

Rohlf FJ, Marcus LF (1993) A revolution morphometrics. Trends Ecol Evol 8:129–132

Ruttner F, Tassencourt L, Louveaux J (1978) Biometrical statistical analysis of the geographic variability of *Apis mellifera* L. Apidologie 9:363–381

Shingleton AW, Das J, Vinicius L, Stern DL (2005) The temporal requirements for insulin signaling duringdevelopment in Drosophila. PLOS Biol 3:e289

Smith DR, Slaymaker A, Palmer A, Kaftanoglu O (1997) Turkish honey bees belong to the east Mediterranean mitochondrial lineage. Apidologie 28(5):269–274

Solignac M, Vautrin D, Loiseau A, Mougel F, Baudry E, Estoup A, Garnery L, Haberl M, Cornuet JM (2003) Five hundred and fifty microsatellite markers for the study of the honey bee (*Apis mellifera* L.) genome. Mol Ecol Notes 3(2):307–311

Spötter A, Gupta P, Nürnberg G, Reinsch N, Bienefeld K (2012) Development of a 44K SNP assay focussing on the analysis of a varroa-specific defence behaviour in honey bees (*Apis mellifera carnica*). Mol Ecol Resour. https://doi.org/10.1111/j.1755-0998.2011.03106.x

Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS et al (2006) Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. Science. https://doi.org/10.1126/science.1132772

Yeh FC, Yang RC, Boyle TBJ, Ye ZH, Mao JX (1997) POPGENE, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Canada

Yıldız Bİ, Tüten E, Aydın S, Aslan YK, Çetin R, Sur E, Karabağ K (2023) A study of whether the genetic variation decreased or not in the protected Caucasian bee, *Apis mellifera caucasica* Pollmann, 1889 (Hymenoptera: Apidae) population in isolated regions. Turk J Entomol. https://doi.org/10.16970/entoted.1273612

Zelditch ML, Swiderski DL, Sheets HD (2012) Geometric morphometrics for biologists: a primer. Elsevier Science Publishing, San Diego

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.