Automated Identification of *Ectatomma edentatum* (Hymenoptera: Formicidae) using Supervised Algorithms

Amanda Araujo de Jesus Santos[1*], Julio Oliveira Silva[1], Deise Machado Lima[1], Vagner Viana Araujo[1], Jacques Hubert Charles Delabie[1], Eltamara Souza Conceição[1]
[1]UNEB, PPGMSB; Alagoinhas, Bahia, Brazil

Taxonomy constantly seeks alternatives to simplify and enhance the identification of living organisms. This study focuses on developing new tools for identifying ant species, aiming to address gaps in determining certain species that pose challenges for naming and study. Species identification can often be a time-consuming and intricate process. We aim to automate the identification process of *Ectatomma* edentatum (Roger, 1863), utilizing Machine Learning techniques to assess if efficiency can be improved and gaps in ant taxonomy reduced. We applied k-nearest neighbors (KNN) and Support Vector Classification (SVC) algorithms. Adapting these models to the dataset yielded excellent results, with both models demonstrating positive performance in classifying the ants. SVC achieved 100% accuracy, while KNN achieved 96%, affirming the effectiveness of these methods in ant identification. This study highlights the value of supervised algorithms in myrmecology, offering a valuable tool for taxonomy and species classification, ultimately providing accurate synthesis and prediction for species naming.
Keywords: Formicidae. *Ectatomma*. Machine Learning.

Introduction

Contextualization and Relevance of Identifying Ant Species

Identifying ant species is pivotal for assessing and synthesizing geographic and ecological data globally. This process enables researchers to investigate the impacts of environmental changes, such as habitat fragmentation and climate alterations, on ant populations and the ecosystems they inhabit [1-3]. By accurately identifying Formicidae, scientists can gain insights into these insects' complex interactions with the environment and humans.

Challenges Inherent in Manual Identification and the Need for Automated Methods

Manual identification of ant species poses significant challenges due to their high diversity and difficulty distinguishing between genera and species. Machine learning and artificial intelligence (AI) techniques can address these challenges to improve the analysis and identification of patterns, serving as supportive tools to facilitate manual execution [4,5]. Given the importance of ants to ecosystems and their widespread existence in various habitats worldwide, using these algorithms for species identification becomes viable and essential [6,7].

Genus *Ectatomma*, the Study's Objective, and an Overview of the Proposed Approach

The genus *Ectatomma*, found in Neotropical and Nearctic regions, comprises giant ants characterized as generalist, polyphagous predators with epigeal and hypogeal habits [8-12]. Recent phylogenomic studies using molecular markers have enhanced our understanding of the evolutionary relationships within the genus *Ectatomma* and between its subfamilies Ectatomminae and Heteroponerinae [13]. This research has led to a new classification for the subfamilies and the description of a new genus.Additionally, studies have highlighted the importance of biogeography and cryptic diversity

Received on 21 Novermber 2023; revised 17 December 2023.
Address for correspondence: Amanda Araujo de Jesus Santos. BR 110, Km 03, Alagoinhas. Zipcode: 48.000.000. Alagoinhas, Bahia, Brazil. E-mail: amandappgmsb@gmail.com.

J Bioeng. Tech. Health 2023;6(Suppl2):1-8

in comprehending the evolution of this ant group [14]. Consequently, this study uses machine learning to automate the identification process of ant species belonging to the genus *Ectatomma*. The objective is to enhance the efficiency of identification processes and reduce gaps in ant taxonomy by leveraging machine learning techniques.

Materials and Methods

Overview of Machine Learning's Field and Its Application in Data Classification

The biological material utilized in this study originated from the Zoology Laboratory collection at the State University of Bahia (UNEB) Campus II, located in Alagoinhas-BA. These specimens were obtained from studies conducted in Bahia's Northern and Agreste Coastal Identity Territory between 2011 and 2023. Thirty specimens of *Ectatomma edentatum* (Roger, 1863) were selected for testing purposes.

Parameters were chosen based on established guidelines to measure the morphological characteristics of the ants [15]. These parameters included antenna, gaster (dorsal, lateral), mesosoma (dorsal, lateral), head length and width, interocular distance, eyes, petiole (dorsal, lateral), femur, and tibia. Measurements were conducted using a stereomicroscope with an HD LITE 1080P camera and Capture 2.3 software for image capture and measurement. The programming language employed for data analysis was PYTHON, utilizing a Jupyter Notebook version 6.4.12.

Selection and Delineation of Classification Algorithms

Previous studies in automated ant classification predominantly focused on image recognition using convolutional neural networks (CNN). As there were no examples of ant classification using supervised machine learning algorithms, the models for this study were selected randomly. The models were k-nearest neighbors (KNN) and Support Vector Classification (SVC). These algorithms were implemented in their standard versions without any calibration or alterations to internal parameters.

The dataset was divided into two sets: 70% for training and 30% for testing to evaluate the performance of the models. This division ensured that the models were trained on sufficient data while allowing for unbiased testing on unseen data.

Results and Discussion

Presentation of Results Obtained through the Application of Classification Algorithms

Both k-nearest neighbors (KNN) and Support Vector Classification (SVC) models demonstrated excellent adaptation to the dataset. KNN achieved a desirable 96% accuracy, while SVC exhibited a perfect 100% accuracy (Table 1). This performance underscores the effectiveness of these methods in ant classification.

The performance of the classification models in the tested scenario was excellent, as evidenced by their high accuracy rates. Moreover, personalized hyperparameters in the prediction process exhibited satisfactory results, with improvements observed in 68 out of 120 performance metrics, indicating the effectiveness of these customized settings. In a separate study involving the classification of Primary Progressive Aphasia (PPA) types based on performance in the TROG-Br Grammar Reception Test, the Decision Tree and Support Vector Classification (SVM) algorithms outperformed Naive Bayes and k-nearest neighbors (KNN)

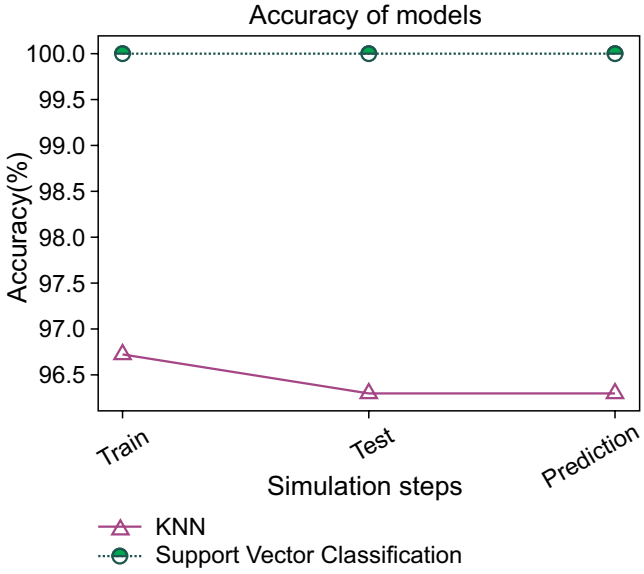
Table 1. Training, testing, and predictive reference values for supervised algorithms.

Model	Train	Test	Prediction
KNN	96.72	96.29	96.29
SVC	100.0	100.0	100.0

methods, highlighting the superiority of these algorithms in specific contexts [16]. Similarly, in an analysis of yellow fever occurrence in Minas Gerais, the Random Forest model demonstrated superior performance to other machine learning models such as Xgboost, SVM, and a neural network [17].

The behavior of the algorithms, when applied to the ant dataset underscores their robustness and reliability (Figure 1), particularly in terms of accuracy. This performance evaluation further proves their effectiveness in accurately classifying ant species. Overall, these findings emphasize the versatility and efficacy of machine learning algorithms, demonstrating their ability to excel across various domains and datasets, including the classification of ant species.

Figure 1. Simulation of the accuracy of supervised models.



However, there are other ways to measure the efficiency and assertiveness of a classification algorithm.

In addition to evaluating the accuracy, a comprehensive analysis of classification metrics was conducted, including precision, recall, F1 score, and ROC score, to provide a more thorough assessment of algorithm performance (Table 2). Consistently high performance across these metrics further confirms the efficiency of the algorithms in accurately classifying ant species. When comparing these results to previous studies, particularly those by Freitas and colleagues [16], where machine learning algorithms demonstrated evaluation performance below 70% in identifying Primary Progressive Aphasia (PPA) using the confusion matrix, the contrast highlights the robustness and effectiveness of the algorithms employed in our study.

Furthermore, in a related study by Wang and colleagues [18], a new automated system for identifying insect images achieved high accuracy rates, with the system attaining 93% accuracy when testing nine regular orders and suborders of insects using artificial neural networks (ANNs) and support vector machines (SVM). This underscores the potential effectiveness of machine learning approaches, particularly SVM, in accurately classifying diverse biological specimens.

Analysis of the Most Relevant Characteristics for the Accurate Identification of Ant Species

Despite the excellent overall results, a detailed analysis was conducted to identify the most relevant attributes or characteristics for accurate ant species identification. This investigation aimed to optimize the machine learning process and facilitate data visualization (Figure 2).

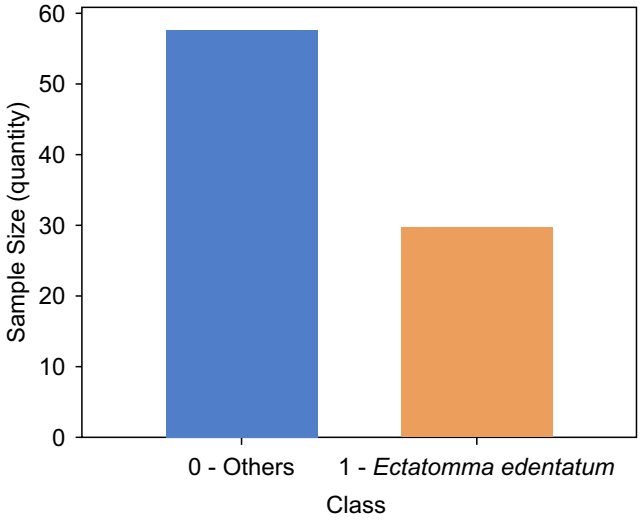
In constructing the dataset, we focused on *Ectatomma edentatum* as the target species and included two other frequently encountered

Table 2. Algorithm evaluation metrics.

Model	Accuracy	Precision	Recall	F1	Roc_auc
KNN	96.2	96.4	87.5	93.3	93.7
SVC	100.0	100.0	100.0	100.0	100.0

species in the region, *Ectatomma* opaciventre and *Ectatomma tuberculatum*, for comparison. As depicted in Figure 2, these species were categorized into two groups based on the number of specimens analyzed, with *E. edentatum* forming one group and the other two forming another, as expected.

Figure 2. Distribution of ants by categories.



Two methods were employed to determine the optimal number of attributes for investigation:

1) Recursive Feature Elimination with cross-validation (RFECV) using Linear Regression; and
2) Recursive Feature Elimination (RFE) using Linear SVC.

Both methods indicated that seven features out of the 14 attributes in the dataset were ideal (Figure 3).

This finding is precious for facilitating manual taxonomy work, suggesting that measuring all 14 parameters for each species may be optional. Instead, focusing on seven key features can yield optimal results. The correlation matrix was then generated, with correlations between paired and stacked features, enabling the identification of the seven traits with the highest pairwise pivot value (Figure 4).

The identified traits include interocular distance, head width, mesodorsal gaster, dorsal gaster, mesolateral gaster, antenna, and head length. Researchers aiming to replicate this study can streamline their data collection process by focusing on measuring these seven ant parameters, leveraging the seven algorithms presented here for accurate identification.

Figure 3. Feature selection using RFE.

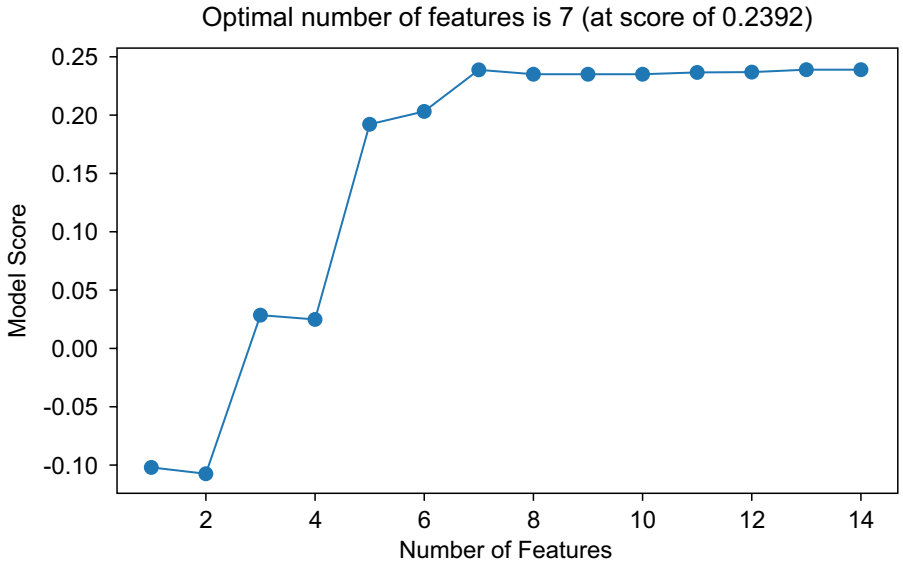
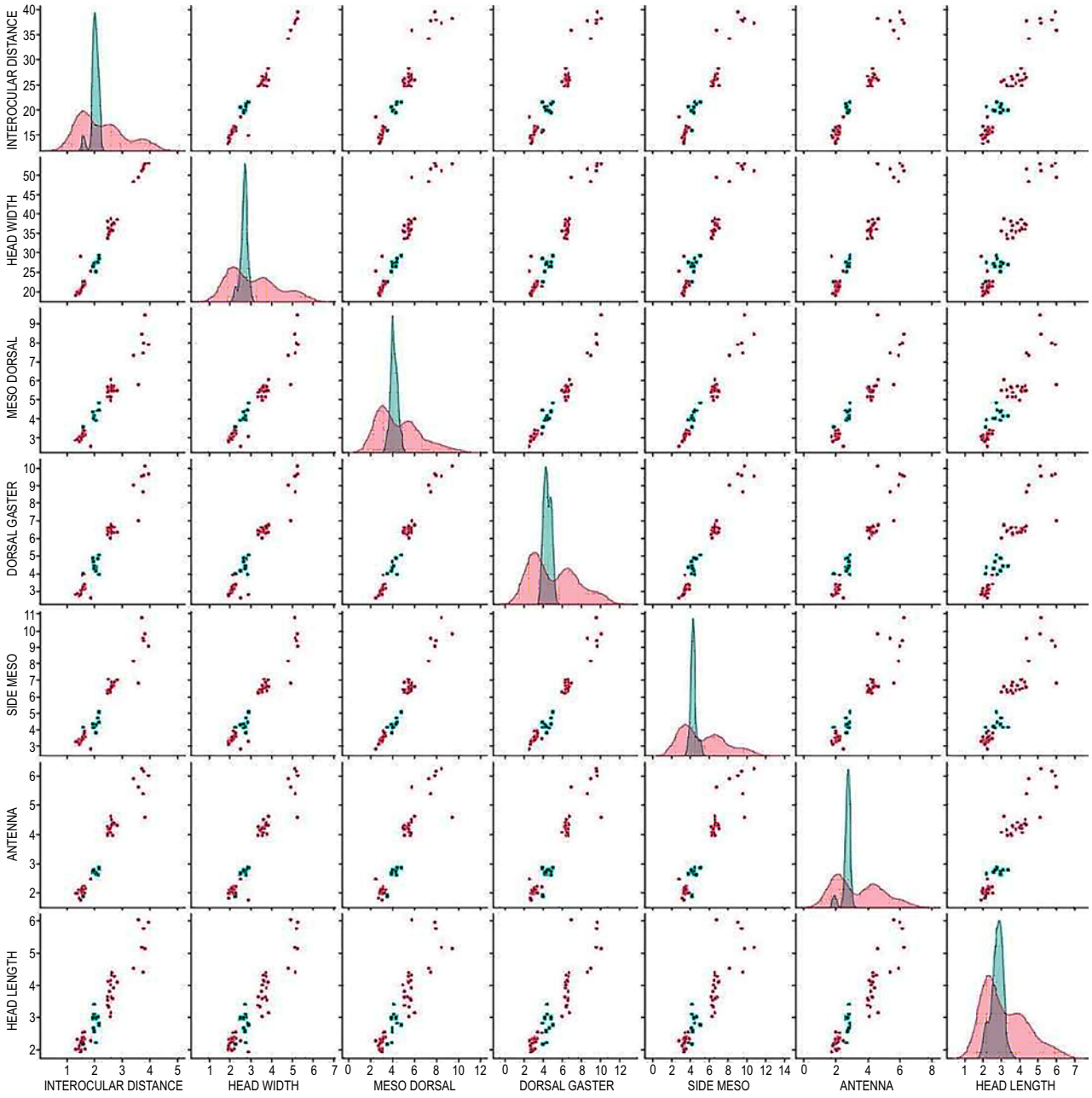


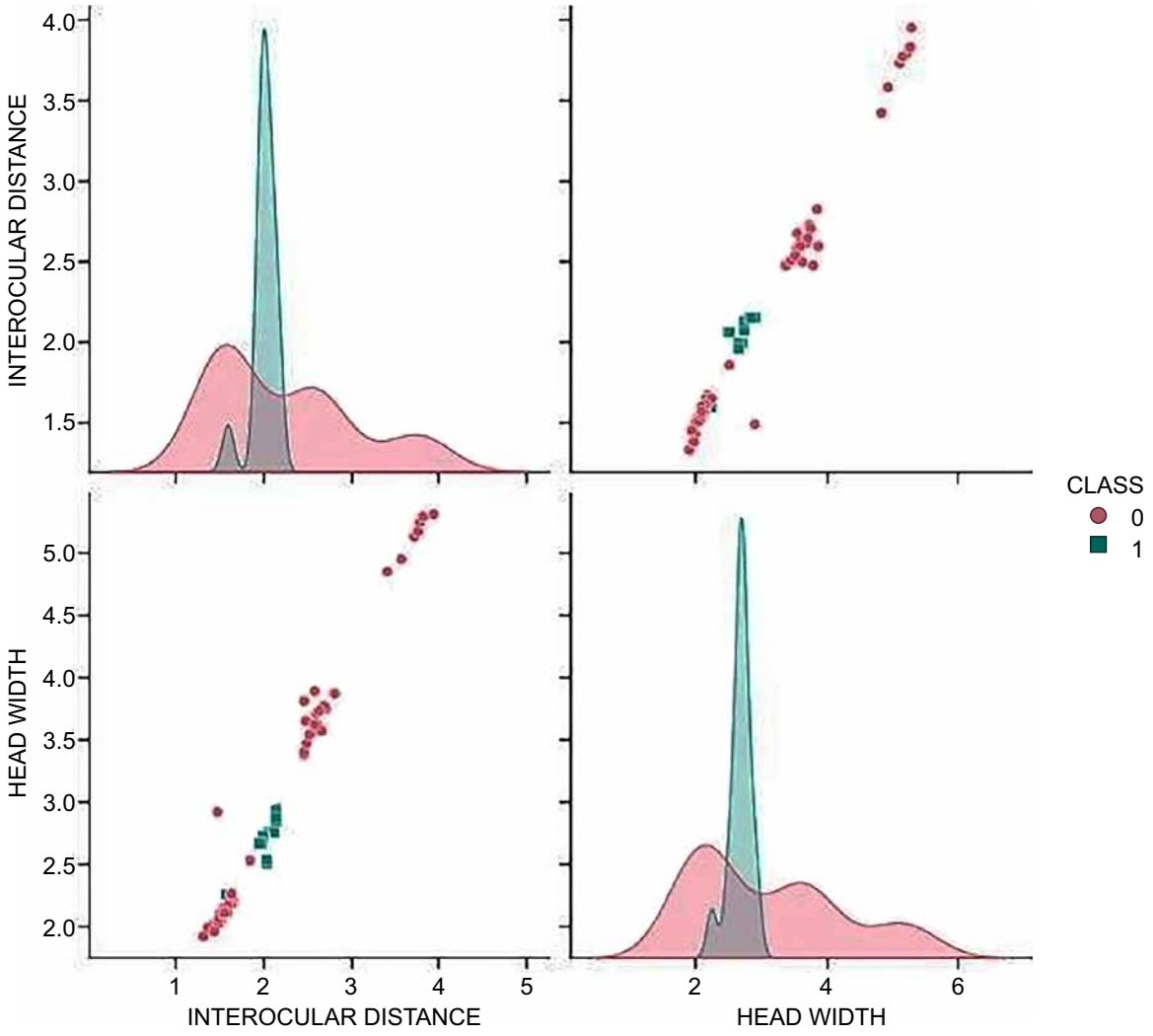
Figure 4. Correlation matrix of the 7 attributes with the highest reference value.



Indeed, with seven attributes, visualizing simultaneously all of them can be challenging. Figure 5 was reconstructed to include only the first three most correlated attributes for a more focused analysis of their cause-and-effect relationships.

The revised visualization shows that the measurement distributions are well-grouped, indicating solid correlations among these variables. This further validates the categorization of these attributes within the group of best features for accurate identification. The focused analysis provided by this visualization enhances our understanding of how these key attributes interact and contribute to the identification process, reinforcing their importance in classifying ant species.

Figure 5. Correlation matrix using 3 attributes that were most correlated in the matrix.



The 3D version of the visualization was constructed to address situations where values overlap, which could distort their interpretation. This interactive approach allows for rotation on the axes and viewing from various angles, providing a clearer understanding of the relationships among the attributes.

Upon interaction with the 3D visualization (Figure 6), it became apparent that there was only a single case of overlap for the three selected features. Significantly, this overlap falls within the error rates of the classification models, indicating that it does not significantly impact the overall accuracy and reliability of the classification process.

The observation that the best parameters produced better results than the entire dataset (Table 3) is noteworthy. This improvement can be attributed to carefully selecting the most correlated and relevant parameters, indicating that these attributes play a crucial role in enhancing the performance of the algorithms. By choosing and focusing on the best parameters, the algorithms achieved higher accuracy and efficiency in the classification process. This underscores the importance of feature selection and highlights how identifying and prioritizing the most relevant attributes can significantly impact the overall performance of machine learning algorithms.

Figure 6. 3D view of the three most related attributes.

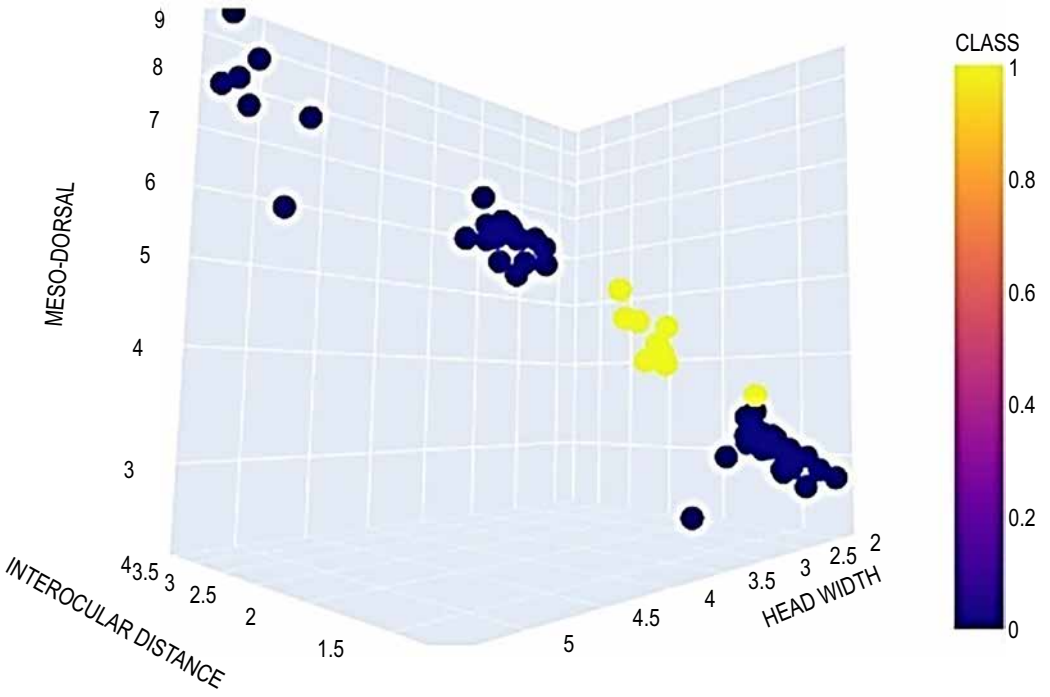


Table 3. Algorithm evaluation metrics using seven attributes.

Model	Train	Test	Prediction
KNN	100.00	100.00	100.0
SVC	100.0	100.0	100.0

Conclusion

In conclusion, the methods employed in this study demonstrated excellent adaptation to the ant identification process, achieving 100% accuracy. This underscores the effectiveness of supervised machine learning algorithms in facilitating the identification of ants. The results affirm the value of these techniques in scientific research, showcasing their ability to synthesize information and accurately predict the species of ants under analysis. Overall, the study highlights the potential of supervised algorithms as valuable tools in taxonomy and classification tasks within myrmecology.

References

1. Gibb H et al. A global database of ant species abundances. 2017.
2. Arnan X, Cerdá X, Retana J. Relationships among taxonomic, functional, and phylogenetic ant diversity across the biogeographic regions of Europe. Ecography 2017;40(3):448-457.
3. Klink RV et al. InsectChange: a global database of temporal changes in insect and arachnid assemblages. Ecology 2021;102(6).
4. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-444.
5. Frid-adar M et al. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 2018;321:321-331.
6. Barros LAC et al. Cytogenetic studies in Trachymyrmex holmgreni Wheeler, 1925 (Formicidae: Myrmicinae) by conventional and molecular methods. Sociobiology 2018;65(2):185-190.
7. Brown JR, WL. Contributions toward a reclassification of the Formicidae. II. Tribe *Ectatommini* (Hymenoptera). Bulletin of the Museum of Comparative Zoology at Harvard College 1958;118:175-362.
8. Kugler C. Brown JR, WL. Revisionary & other studies on the ant genus *Ectatomma*, including the descriptions

of two new species. Search Agriculture- New York State Agricultural Experiment Station, Ithaca, 1982.
9. Achaud JP, Pérez-lachaud G. Ectaheteromorph ants also host highly diverse parasitic communities: a review of parasitoids of the Neotropical genus *Ectatomma*. Insectes Sociaux 2015;62:121-132.
10. Fernández F. Las hormigas cazadoras del género *Ectatomma* (Formicidae: Ponerinae) en Colombia. Caldasia 1991:551-564.
11. Del-Claro K, Oliveira PS. Ant-homoptera interactions in a Neotropical savanna: The honeydew-producing treehopper, *Guayaquila xiphias* (Membracidae), and its associated ant fauna on *Didymopanax vinosum* (Araliaceae) 1. Biotropica 1999;31(1):135-144.
12, Lachaud J, Pérez-Lachaud G, Heraty JM. Parasites associated with the ponerine ant *Ectatomma tuberculatum* (Hymenoptera: Formicidae): first host record for the genus *Dilocantha* (Hymenoptera: Eucharitidae). The Florida Entomologist 1998;81(4):570-574.
13. Camacho GP et al. UCE phylogenomics resolves major relationships among ectaheteromorph ants (Hymenoptera: Formicidae: Ectatomminae, Heteroponerinae): A new classification for the subfamilies and the description of a new genus. Insect Systematics and Diversity 2022;6(1):5.
14. Nettel-Hernanz A et al. Biogeography, cryptic diversity, and queen dimorphism evolution of the Neotropical ant genus *Ectatomma* Smith, 1958 (Formicidae, Ectatomminae). Organisms Diversity & Evolution 2015;15:543-553.
15. Silva-Freitas JM, Mariano CF, Delabie JHC.Morphometry Formicidae. Postgraduate Program in Biological Sciences (Animal Biology). Federal University of Espírito Santo. Vitória, ES, Brazil. Itabuna 2015.
16. Freitas M et al. Uso de aprendizado de máquina para identificar o tipo de afasia progressiva primária a partir do desempenho no Trog-2Br. Anais do Computer on the Beach 2023;14:512-514.
17. Araújo IL et al. Performance comparison of machine learning algorithms for predictive analytics of yellow fever in the state of Minas Gerais. 2023.
18. Wang J et al. A new automatic identification system of insect images at the order level. Knowledge-Based Systems 2012;33:102-110.