Publicación Electrónica
ASOCIACIÓN PALEONTOLÓGICA ARGENTINA

www.peapaleontologica.org.ar

Support vector machine-based classification of Paleozoic gymnosperm woods

MÁRIO G. F. ESPERANÇA JÚNIOR[1]
ROBERTO IANNUZZI[2]
DOMINGAS MARIA DA CONCEIÇÃO[3,4]

1. Universidade Federal do Ceará (UFC), Ceará, Brazil.
2. Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil.
3. Museum of Natural History Chemnitz (MfNC), Chemnitz, Saxony, Germany
4. Universidade Regional do Cariri (URCA), Museu de Paleontologia Plácido Cidade Nuvens, R. Plácido Cidade Nuvens, 326, Santana do Cariri, Ceará, Brazil.

Recibido: 9 de enero 2026 - **Aceptado:** 25 de junio 2026 - **Publicado:** 9 de septiembre 2026

Para citar este artículo: Mário G. F. Esperança Júnior, Roberto Iannuzzi, & Domingas Maria Da Conceição (2026). Support vector machine-based classification of Paleozoic gymnosperm woods. *Publicación Electrónica de la Asociación Paleontológica Argentina* 26(3): 1–14.
Link a este artículo: http://dx.doi.org/10.5710/PEAPA.25.06.2026.562

ISSN 2469-0228

Asociación Paleontológica Argentina
Maipú 645 1° piso, C1006ACG, Buenos Aires
República Argentina
Tel/Fax (54-11) 4326-7563
Web: www.apaleontologica.org.ar

ARTÍCULO Año 2026 - 26(3): 1–14 PeAPA Publicación Electrónica ASOCIACIÓN PALEONTOLÓGICA ARGENTINA

SUPPORT VECTOR MACHINE-BASED CLASSIFICATION OF PALEOZOIC GYMNOSPERM WOODS

MÁRIO G. F. ESPERANÇA JÚNIOR[1], ROBERTO IANNUZZI[2], AND DOMINGAS MARIA DA CONCEIÇÃO[3,4]

[1]Universidade Federal do Ceará (UFC), Ceará, Brazil. *mario.esperancajr@gmail.com*

[2]Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil. *riannuzzi662@gmail.com*

[3]Museum of Natural History Chemnitz (MfNC), Chemnitz, Saxony, Germany. *domingas.paleonto@gmail.com*

[4]Universidade Regional do Cariri (URCA), Museu de Paleontologia Plácido Cidade Nuvens, R. Plácido Cidade Nuvens, 326, Santana do Cariri, Ceará, Brazil.

MGFEJ: https://orcid.org/0000-0002-5338-9000; **RI:** https://orcid.org/0000-0003-1432-8106; **DMDC:** https://orcid.org/0000-0003-2917-0065

Abstract. Understanding the taxonomic placement of Paleozoic gymnosperm woods remains a complex task largely due to the frequent absence of attached reproductive or foliar structures. Machine-assisted analyses based on wood anatomy can support this task by suggesting plausible affinities between fossil genera and taxonomic orders for paleobotanists to further explore. In this context, we developed a support vector classifier employing a soft-margin, which achieved a maximum estimated accuracy of 71.4% and a macro-averaged F1-Score of 61% across four evaluated regularization values, representing an improvement over previously proposed models. The null hypothesis of the permutation test by label shuffling was rejected ($p < 0.001$), indicating that the model effectively learned the underlying structure linking anatomical features to taxonomic orders. To evaluate its applicability, six genera from western Gondwana were selected as input to test the classifier, with predicted affinities either aligning with previous literature or offering new perspectives for future paleobotanical research. The model proposed is data-driven and capable of learning from patterns in anatomical characters, operating through statistical generalization across examples. While the model's main limitations are rooted in biological and taphonomical constraints, its overall performance demonstrates the usefulness of machine learning approaches as an additional aid to fossil wood taxonomy.

Key words. Supervised learning. Wood anatomy. Taxonomy. Computational paleontology. Paleobotany.

Resumen. CLASIFICACIÓN DE MADERAS DE GIMNOSPERMAS PALEOZOICAS BASADA EN MÁQUINAS DE SOPORTE VECTORIAL. Comprender la ubicación taxonómica de las maderas de gimnospermas paleozoicas sigue siendo una tarea compleja, en gran parte debido a la frecuente ausencia de estructuras reproductivas o foliares asociadas. Los análisis asistidos por computadora basados en la anatomía de la madera pueden apoyar esta tarea al sugerir afinidades plausibles entre géneros fósiles y órdenes taxonómicos que los paleobotánicos pueden explorar más a fondo. En este contexto, desarrollamos un clasificador de soporte vectorial utilizando un margen suave, que alcanzó una exactitud estimada máxima del 71,4% y un F1-Score macro-promediado del 61% entre los cuatro valores de regularización evaluados, lo cual representa una mejora respecto a los modelos previamente propuestos. La hipótesis nula de la prueba de permutación mediante el reordenamiento de etiquetas fue rechazada ($p < 0.001$), lo que indica que el modelo efectivamente aprendió la estructura subyacente que vincula las características anatómicas con los órdenes taxonómicos. Para evaluar su aplicabilidad, se seleccionaron seis géneros de Gondwana occidental como entrada para probar el clasificador, cuyas afinidades predichas coincidieron con la literatura previa o bien ofrecieron nuevas perspectivas para futuras investigaciones paleobotánicas. El modelo propuesto está basado en datos y es capaz de aprender a partir de patrones en los caracteres anatómicos, operando mediante generalización estadística a partir de ejemplos. Si bien las principales limitaciones del modelo tienen su origen en restricciones biológicas y tafonómicas, su desempeño general demuestra la utilidad del aprendizaje automático como una ayuda complementaria para la taxonomía de maderas fósiles.

Palabras clave. Aprendizaje supervisado. Anatomía de la madera. Taxonomía. Paleontología computacional. Paleobotánica.

INTRODUCTION

Fossil plants have been applied in numerous studies to understand the evolution of the plant kingdom, interpret ancient environments and climates, determine relationships with other organisms, and aid in the dating and correlation of rocks (Iannuzzi & Vieira, 2005; Taylor et al., 2009). An issue often associated with macrofossils is the disconnection between organs during the life cycle or due to the transport or incorporation of bioclasts, as well as the selective preservation of parts with different compositions and resistance. Consequently, fossil assemblages commonly consist of similar types of remains rather than com-

plete individuals, reflecting both biological and taphonomic biases (Simões et al., 2010). The case of plants is not an exception, highlighting the inherent complexity involved in reconstructing whole organisms and accurately assigning them to taxonomic groups (Greenwood, 1991; Spicer, 1991; Ferguson, 2005). In the study of fossil woods, the fragmentary nature of the material is a recurrent feature throughout the geological record. In most instances, preservation is restricted to fragments of secondary xylem, which considerably constrains taxonomic resolution at higher hierarchical levels (Rößler et al., 2014; Philippe et al., 2019; Conceição et al., 2023). This limitation is particularly evident in Paleozoic plant assemblages from the Gondwanan floristic realm, where specimens are abundant but often have poorly understood taxonomic resolution (e.g., Guerra-Sommer, 1977; Mussa, 1978, 1986a, 1986b; Crisafulli, 1998, 2001; Merlotti, 2002, 2009; Kurzawe et al., 2013; Decombeix et al., 2019; Wan et al., 2020; Conceição et al., 2020a, 2022).

In recent decades, artificial intelligence has gained increasing attention from the paleontological community as a tool for problem-solving and decision-making. While artificial neural networks and deep learning techniques have become more widespread in recent applications, traditional algorithms — such as k-nearest neighbors, random forests, and support vector machines — continue to fulfill a significant role in studies involving fossils (Martín-Perea et al., 2020; Yu et al., 2024; Liu et al., 2026; Silvestro & Pimiento, 2025; Yaqoob et al., 2025; MacFadden et al., 2026). Given the challenges associated with Paleozoic gymnosperm woods, Conceição et al. (2023) proposed a machine learning classifier based on stem anatomy (*Paleo Wood*), offering new perspectives despite the limitations observed in the models. The performance of the three tested algorithms was similar, that is, k-nearest neighbors, logistic regression, and linear discriminant. Reported issues were mainly related to the limited sample size and the high variability within some classes (Conceição et al., 2023). Therefore, to reduce misclassifications in this type of dataset, the selected models should be able to better handle class imbalance and intraclass variability.

In this context, support vector classifiers have shown potential for improving classification accuracy while maintaining model interpretability. Beyond classification, support vector machines have been applied to a wide range of tasks, with increasing adoption in recent decades, including regression, outlier detection, and prediction of complex structures (Tsochantaridis et al., 2005). Notable applications include text categorization, handwritten character recognition, image and audio processing, video analysis, object and face recognition, speaker identification, bioinformatics tasks and medical diagnosis, as well as computer vision and security (Burges, 1998; Hearst et al., 1998; Moguerza & Muñoz, 2006; Lorena & De Carvalho, 2007; Cristianini & Ricci, 2008; Ma & Guo, 2014; Awad & Khanna, 2015; Cervantes et al., 2020). The goal of this learning algorithm is to project the data into a higher-dimensional space, enabling class separation through hyperplanes with margins determined by the greatest possible distance. In practice, this is achieved through kernel functions, which transform the input data while operating in a reduced subspace defined by the support vectors, where the data points that lie closest to the hyperplane contain all the information necessary to reconstruct it (Mammone et al., 2009; James et al., 2013). Such models are supported by a solid theoretical foundation in computational learning theory and, when properly tuned, typically exhibit strong generalization performance while being less sensitive to input biases. They yield a unique global solution, unlike some learning algorithms that may converge to different local optima (Hearst et al.,1998; Burbidge & Buxton, 2001; Karamizadeh et al., 2014).

Building on these insights, we propose herein a support vector classifier, which tends to be more robust for analyzing the wood anatomy of ancient seed plants for their subsequent classification within higher hierarchical categories, as previously explored using other algorithms by Conceição et al. (2023). The main objective is to link Paleozoic gymnosperms woods to their corresponding taxonomic orders, and it is not our goal to automate the classification process. Additionally, we compare this new approach with that previously presented in *PaleoWood* and apply the model to certain genera from the western Gondwanan floristic realm.

MATERIALS AND METHODS

Materials and Data

In order to develop a model with a potentially better performance than *PaleoWood*, we used the anatomical dataset from Paleozoic gymnosperm woods compiled by Conceição et al. (2023). This dataset consists of 16 input characters recorded for 42 plant-wood genera (Supplementary Material, Table S1), each associated with one of 11 taxonomic gymnosperm orders as output. All the selected samples have a complete set of described characters, with no missing values. The input variables are binary or ordinal and represent anatomical features of the inner core of plant, including the presence or absence of tissues and cells, maturation pattern of the primary xylem, characteristics of the rays and pits, as well as the wood type (Tab. 1). Features related to the bark, phyllotaxis or leaf traces were not included, given their infrequent preservation in the fossil record and challenges in reliable detection, although useful in determining higher taxonomic groups (Dunn, 2006). The output classes are categorical and fall within either the pteridospem group — Buteoxylonales, Calamopityales, Callistophytales, Gigantopteridales, Glossopteridales, Lyginopteridales, and Medullosales — or other gymnosperms, named Cordaitales, Cycadales, Ginkgoales, and Voltziales. Output labels are remarkably imbalanced. Further details on the variables are provided by Conceição et al. (2023).

Additionally, six other genera from western Gondwana were set aside for application of the final model. These include *Amosioxylon* Césari et al., 2005, *Eoguptioxylon* Crisafulli et Lutz, 2007, *Iratinia* Spiekermann et al., 2021, *Parnaiboxylon* Kurzawe et Merlotti, 2013, *Paulistoxylon* Mussa, 1986, and *Teresinoxylon* Mussa, 1989 (see Mussa, 1986a; Caldas et al., 1989; Césari et al., 2005; Crisafulli & Lutz, 2007; Kurzawe et al., 2013; Spiekermann et al., 2021). To classify these genera and evaluate model performance,

TABLE 1 – Input variables as in Table S1 (Supplementary Material), according to Conceição et al. (2023).

Variable	Type	Meaning
Tracheid	Binary	Presence of tracheids in the central core
Parench	Binary	Presence of parenchyma in the central core
Resin	Binary	Presence of resiniferous cells, channels or ducts in the central core
Scleren	Binary	Presence of sclerenchyma in the central core
Septa	Binary	Presence of septations in the central core
Trilob	Binary	Presence of trilobate or triribbed stele or petioles
Endarch	Binary	Presence of endarch maturation of the primary xylem
Mesarch	Binary	Presence of mesarch maturation of the primary xylem
Exarch	Binary	Presence of exarch maturation of the primary xylem
MixPit	Binary	Presence of mixed pits on the radial walls of tracheids
PitSeriat	Ordinal	Seriation of pits on the radial walls of tracheids
TangPit	Binary	Presence of pits on the tangential walls of tracheids
AxialPar	Binary	Presence of axial parenchyma in the secondary xylem
RayWidth	Ordinal	Width of rays
RayHeight	Ordinal	Average rays height
WoodType	Ordinal	Type of wood

given the limited sample size and the high-dimensional nature of the dataset, we implemented a support vector machine approach. Prior to the model training, a data processing step was applied to emphasize anatomical traits considered particularly informative for classification purposes.

Data Processing. To account for the relative importance of certain structural attributes, a weight of 2 was assigned to the variable Tracheid, which indicates the presence or absence of these cells in the central core of stems. This decision was based on the relevance of this feature to the evolutionary development of the axis (Beck et al., 1982) and, on the observation that models tend to underestimate its contribution during training, as previously reported by Conceição et al. (2023). The increased weight was implemented during model fitting, effectively amplifying the contribution of this variable in the decision function. All other input features were assigned a weight of 1. Additionally, feature scaling was not applied, so no normalization or standardization was performed on the input data. This choice aimed to preserve the explainability of the original variable encodings, which are binary or ordinal in nature. These adjustments aimed to enhance the sensitivity of the classifier to biologically meaningful distinctions, while preserving overall model interpretability. The processed dataset was then used to train the support vector models analyzed in the following section, with the methodological framework and theoretical formulation adopted are detailed below.

Methods

Support Vector Machines Principles and Formulation. Support vector machines (SVM) classifiers aim to optimize the following primal formulation (Eq. 1), minimizing the norm of the weight vector w, while penalizing misclassifications through slack variables $\xi_i \geq 0$, scaled by the regularization term C. The vector w is orthogonal to the optimal decision hyperplane, defining its orientation and thus contributing to the distinction of the n training samples into two classes $y \in \{-1, +1\}$, based on the input x and a bias b (Eq. 1). The inner product between the weight vector and the input is used directly herein, with no data transformation (Eq. 1), that is the linear kernel. Accordingly, this kernel represents the simplest formulation with no feature mapping induced, being the model fitted in the original feature space, a technique often appropriate and computationally efficient in small-sample, high-dimensional settings (Hsu et al., 2003). These expressions represent the soft-margin version of the linear SVM algorithm (Cortes & Vapnik, 1995), in which lower values of C favor model generalization at the cost of increased classification error, whereas higher values of C result in improved fit to the training data but may lead to overfitting.

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \text{ subject to } y_i\left(w^{\mathrm{T}}x_i + b\right) \geq 1 - \xi_i \quad (1)$$

Notably, data that effectively contribute to model training are associated with non-zero slack variables ξ_i, which is the case for points that violate the margins lying close to the decision hyperplane, including misclassified samples — the so-called support vectors. This is a key advantage of this learning algorithm as it fits the model based primarily on the samples that require more attention.

Although the primal formulation of SVM consists of minimizing the norm of the weight vector with penalties (Eq. 1), the problem is typically numerically solved using its dual formulation involving the Lagrangian multipliers α (Eq. 2) and performed via the Sequential Minimal Optimization algorithm (see Platt, 1998). In this study, the implementation was carried out using the e1071 package (Meyer et al., 2024), which serves as an interface to the LIBSVM library (Chang & Lin, 2011). This solution is subject to constraints on the optimization variables, maintaining $C \geq \alpha \geq 0$ and ensuring that the sum of the Lagrangians is equal between both classes.

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i \cdot x_j\right) \quad (2)$$

In this sense, the Lagrangian multipliers can be interpreted as weights applied to the training set, being zero for samples that do not affect the decision boundary, and positive for the support vectors. Therefore, the vector w and the bias term b can be estimated, defining the decision function for the model. The weight vector w is obtained by summing the product of the Lagrangian multiplier α, the class label y, and the input vector x for all support vectors.

The bias term b is typically estimated by solving it in the constraint of Eq. 1, assuming equality and $\xi_i = 0$, and then averaging the result over all support vectors. The decision function is based on the relative position with respect to the hyperplane $w \cdot x_i + b$; that is, a positive result classifies a new sample as belonging to class +1 while a negative result assigns it to class -1 and, a zero value would indicate a sample lying exactly on the hyperplane. While this formulation applies to binary classification, in fact, it must be emphasized that many real-world problems involve more than two classes.

The present study addresses such a multiclass scenario. In this case, the one-against-one strategy was adopted, in which hyperplanes are constructed for each pair of classes, excluding the others. The final classification is determined by a voting mechanism involving all possible pairs, with the class receiving the majority of votes being selected.

Validation Strategy and Comparative Analysis. The SVM model's predictive performance was evaluated analytically through confusion matrices, by comparing the actual and predicted classes for four values of the regularization parameter C: 0.1, 1, 10, and 100. The values on the main diagonal of the confusion matrix represent the correct classifications (true positives), while results for each class located outside the main diagonal correspond to false positives (in the columns) or false negatives (in the rows). The generalizability of these models was assessed through leave-one-out cross validation (LOOCV), a technique in which a single sample is removed from the training set and then used as a test input. After each iteration, the sample is returned to the training set, and the procedure is repeated until every sample has been tested exactly once (James et al., 2013). The adopted strategy is justified by the small sample size relative to the number of classes, a common challenge when dealing with macrofossils (Yu et al., 2024).

Accuracy and macro-averaged F1-Score were estimated based on results obtained through LOOCV. Accuracy provides an overall metric for each model and is calculated as the ratio of true positives to the total number of predictions. In contrast, F1-Score is the harmonic mean between precision and recall for each class individually; precision being the proportion of true positives among all predicted positives (including false positives), and recall being the proportion of true positives among all actual positives (including false negatives). The macro-averaged F1-Score corresponds to the unweighted average across all classes, thus offering a fairer evaluation in the presence of class imbalance (Sokolova & Lapalme, 2009).

Additionally, a permutation test was conducted to assess whether the performance of the SVM classifier was statistically significant, that is, to determine whether the association between classes and features in the dataset reflects a true relationship or could have arisen by chance. The test was performed by comparing the original model accuracy with those obtained under label shuffling. In each iteration, the class labels were randomly permuted across all samples, followed by model training and evaluation using LOOCV to generate a new accuracy value. This procedure was repeated 1000 times, resulting in an empirical distribution of accuracies and an empirical p-value. The null hypothesis, that class labels are unrelated to the feature set, was rejected if the original accuracy exceeded the vast majority of the permuted values, indicating that the observed performance is unlikely to have occurred by chance. This suggests that the model successfully learned the underlying structure that relates the data to their corresponding classes (Ojala & Garriga, 2010).

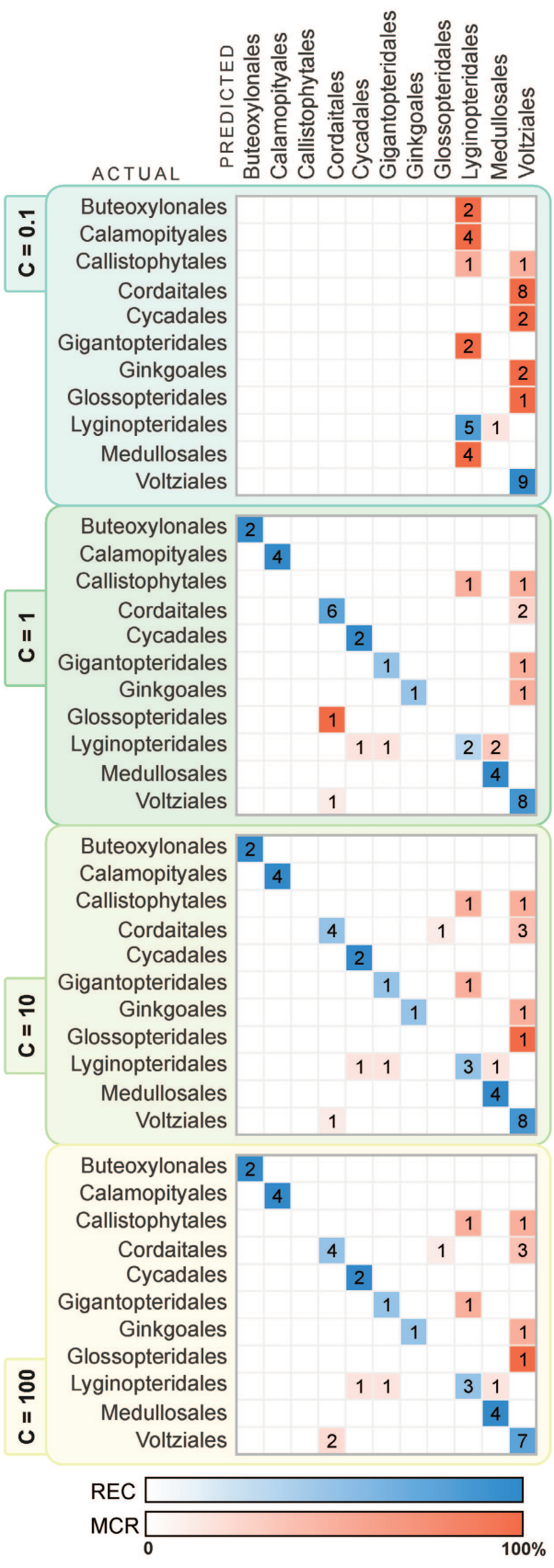
All analyses were performed using RStudio.Version 2024.9.1.394 (Posit Team, 2024). The R script and additional data used in our analysis can be accessed at https://github.com/marioesperancajr/SVMwood. Additional accuracy values for models using more complex kernels are available in Appendix 1.

RESULTS

Model Analysis and Comparison

Through the methodology described, the SVM classifiers were trained using the anatomical dataset compiled by Conceição et al. (2023) to associate Paleozoic axes with gymnosperm orders. Four values of the regularization parameter C were tested and validated using leave-one-out cross-validation (LOOCV), with the results summarized in the confusion matrices (Fig. 1).

When C = 0.1, the model was clearly underfitted, as predictions were concentrated primarily in two groups: Lyginopteridales and Voltziales (Fig. 1). Although this model



is unsuitable for classification purposes, its predictions appear to reflect a broader division into two wood types, corresponding to distinct anatomical features observed among manoxylic pteridosperms and basal conifers or pycnoxylic seed ferns. This pattern had already been identified by Conceição et al. (2023) through dimensionality reduction techniques and has also been noted by several anatomists (e.g., Lacey, 1953; Galtier & Meyer-Berthaud, 2006; Decombeix et al., 2011). Thus, despite its low classification performance, the model demonstrated a certain ability to generalize broader anatomical groupings, even though it failed to distinguish among the specific taxonomic orders within each group.

Using higher values of the regularization parameter led to more realistic predictions (Fig. 1). Misclassifications tended to occur between classes with similar anatomical characteristics, for instance, Cordaitales and Voltziales or Lyginopteridales and Medullosales, as also observed in the *PaleoWood* study (Conceição et al., 2023). Even so, the model's performance improved substantially with SVM, reaching an accuracy of 71.4% for C = 1, in contrast to a maximum of 61.9% achieved using the linear discriminant algorithm (Tab. 2). When analyzing performance more equitably across classes, the highest macro-averaged F1-Score was obtained with C = 10, approximately 61%, compared to a maximum of 55.3% for logistic regression (Tab. 2).

On the other hand, for C = 100, the highest regularization value tested, there is some evidence of overfitting, as the performance metrics did not improve compared to other SVM models (Tab. 2). From the perspective of the primal formulation (Eq. 1), a high regularization parameter imposes a greater penalty on slack variables, severely punishing misclassifications and margin violations. In the dual formulation (Eq. 2), this corresponds to a weaker constraint on the Lagrangian multipliers. As a result, the model begins to learn beyond its true capacity, partially fitting to noise and reducing its generalization ability.

Figure 1. Heatmaps of confusion matrices for the four regularization tested parameters. Values represent the number of classifications obtained through LOOCV. Color intensity corresponds to the proportion of instances, with darker shades indicating higher values of class-wise recall (REC) and pairwise misclassification rates (MCR).

TABLE 2 – Accuracy and macro-averaged F1-Score for different models, including the SVM developed in this study and the *PaleoWood* models[1].

Algorithm	Accuracy	F1-Score (Macro)
SVM (C = 1)	0.714	0.608
SVM (C = 10)	0.690	0.610
SVM (C = 100)	0.667	0.601
LDA	0.619	0.501
LR	0.547	0.553
KNN	0.476	0.440
SVM (C = 0.1)	0.333	0.089

[1]LDA = linear discriminant analysis; LR = logistic regression; KNN = k-nearest neighbors.

The performance of SVM-based classifiers can be sensitive to class imbalance, potentially disadvantaging underrepresented groups by shifting the hyperplanes toward the minority classes, thereby expanding the decision region of the majority class and increasing the rate of misclassification (Akbani et al., 2004; Batuwita & Palade, 2013). This is particularly significant when class imbalance alters the set of support vectors, thereby influencing the slack variables (Eq. 1) and the Lagrange multipliers (Eq. 2). However, the cross-validation results indicate that this effect is not pronounced in the models developed here (Fig. 1). Two minority groups, each represented by at most two samples, did not exhibit false negatives (Buteoxylonales and Cycadales). The single misclassification observed for Glossopteridales should be interpreted with caution due to limitations of LOOCV, as the class is not represented in the training set when that sample is left out. Consequently, the accuracy for minority classes is 0.6 for $C \geq 1$, which is comparable to the overall model performance (Tab. 2). Given that several of these groups exhibit clear anatomical distinctions (Conceição et al., 2023), the impact of class imbalance is likely mitigated by good separability.

To further verify whether the model's performance was meaningful or the result of random associations, a permutation test was conducted by shuffling the class labels and cross-validating the resulting models (Fig. 2). Accuracies equal to or greater than the original were obtained in only 1% of iterations when C = 0.1; and for higher values of C, no permuted accuracy surpassed or matched the original result ($p < 0.001$). Therefore, the null hypothesis of the permutation test is rejected, providing strong evidence that the model effectively captures the underlying data structure associated with each class and

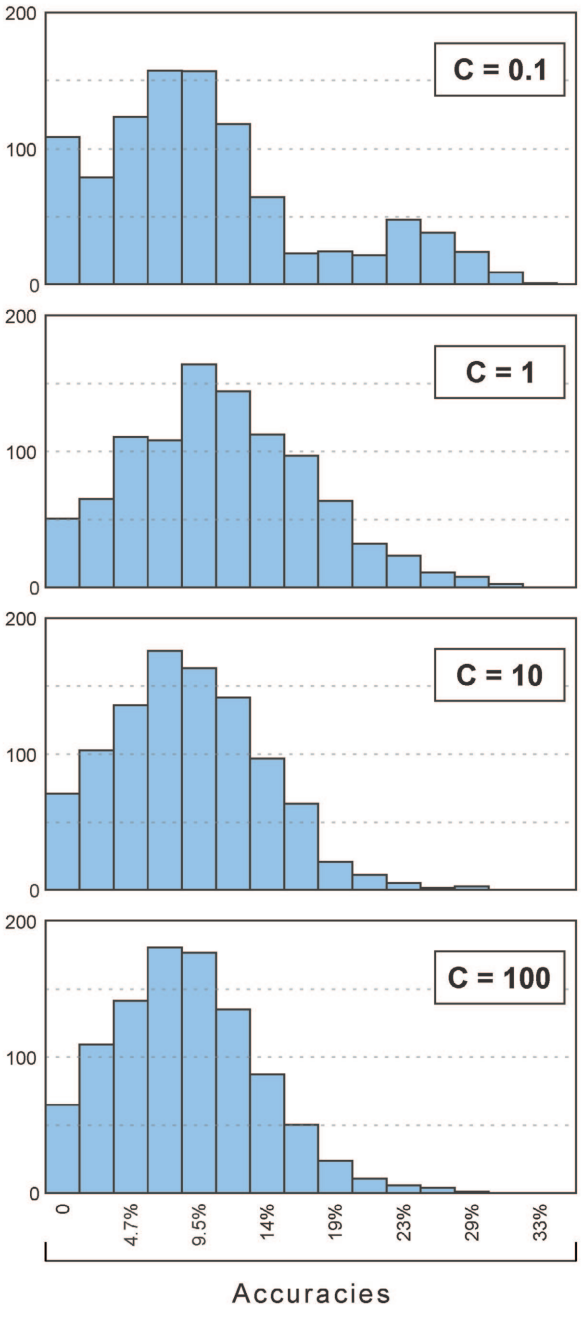


Figure 2. Histograms resulting from the label-shuffling test. Bar length indicate the frequency of cross-validated accuracies across 1000 iterations for each model.

TABLE 3 – Predictions generated by the SVM model for six genera from the western Gondwanan floristic realm.

Genus	Prediction for C = 1	Prediction for C = 10	Affinities previously suggested
Amosioxylon	Medullosales	Medullosales	Calamopityales, Medullosales, Progymnospermopsida (César et al., 2005)
Eoguptioxylon	Cordaitales	Voltziales	Medullosales, Pentoxylales (Crisafulli & Lutz, 2007)
Iratinia	Cycadales	Cycadales	Cycadales (Spiekermann et al., 2021)
Parnaiboxylon	Voltziales	Cordaitales	?
Paulistoxylon	Cordaitales	Voltziales	Cordaitales (Crisafulli, 1998)
Teresinoxylon	Lyginopteridales	Lyginopteridales	Cycadales, Lyginopteridales (Caldas et al., 1989)

that the true positives are not due to random chance. These results support the robustness of the proposed SVM approach, confirming its reliability in classifying Paleozoic gymnosperm woods based on anatomical features. Given these promising outcomes, we proceeded to apply the trained model to six additional genera from the western Gondwanan floristic realm, which were not included in the initial training dataset, using the configurations for which there was no evidence of underfitting or overfitting.

Model application

Comparing predictions from the most suitable models (Tab. 3), three genera yielded consistent results (*Amosioxylon*, *Iratinia*, and *Teresinoxylon*), while the remaining three did not show full agreement among predictions (*Eoguptioxylon*, *Parnaiboxylon*, and *Paulistoxylon*). This variability underscores the importance of analyzing the implications of these predictions in the broader paleobotanical context.

DISCUSSION

Implications of Predictions

The results obtained from the predictions (Tab. 3) can be considered positive, as they align with previous anatomical interpretations. Starting with the genus *Amosioxylon*, assigned to a stem from the Pennsylvanian of Argentina, it displays several vascular strands surrounded by ribbed centrifugal primary xylem and pycnoxylic wood (César et al., 2005). The authors suggested affinities with the Devonian progymnospem *Callixylon*, due to its pycnoxylic wood, as well as with the seed fern *Bostonia* (Calamopityales), based on similarities in their central axis, also recognizing a resemblance to the anatomy of *Medullosa* (Medullosales). These overlapping anatomical traits underscore the taxonomic complexity of *Amosioxylon*, which exhibits features shared across distinct lineages. Nonetheless, the model's prediction aligns with one plausible hypothesis, that is, a link to Medullosales (Tab. 3), which is particularly supported by the presence of medullosan-like central vascular segments. This is a strongly characteristic feature of this group, thus reinforcing the potential of SVM-based approaches to support paleobotanical interpretations.

Eoguptioxylon is a trunk from the late Permian of Argentina that exhibits an unusual combination of features. It presents a polystelic wood, with steles surrounded by sclerenchyma and irregularly arranged in ground parenchyma, in addition to a central diaphragm and pycnoxylic xylem (Crisafulli & Lutz, 2007). The authors suggested a possible affinity with the Paleozoic Medullosales or even the Mesozoic Pentoxylales. Classifications from *PaleoWood* indicated affinities with Cordaitales or Cycadales (Conceição et al., 2023), while the SVM model suggested either Cordaitales or Voltziales (Tab. 3). Although Cordaitales appears as the prevailing assignment, attributing a definitive taxonomic group to *Eoguptioxylon* remains challenging, given the lack of similar stems for confident comparison. This issue will likely remain unresolved until further knowledge on the anatomy of ancient seed plants becomes available.

Iratinia is an anatomically preserved axis with affinities to Cycadales, originating from the early Permian of southern Brazil (Spiekermann et al., 2021). This rare specimen preserves vascular cambium, phloem, cortex, and leaf bases, providing robust evidence for its connection to cycads. Remarkably, even when considering only the anatomical characters of the pith, besides primary and secondary xylem, the SVM model correctly predicted its taxonomic order (Tab. 3), demonstrating its effectiveness in generalizing from limited input data.

Parnaiboxylon currently includes three known specimens: *Parnaiboxylon* sp. 1 and *P. rohnae* from the early Permian of northern Brazil, and *P. wangi* from the Pennsylvanian of China (Kurzawe et al., 2013; Wang et al., 2023). The main features of the genus are the endarch to mesarch primary xylem and the solenoid pith. The computational models identified this taxon as belonging to either Voltziales or Cordaitales (Tab. 3). These non-consensual predictions reflect the scarcity of voltzialean woods exhibiting mesarch primary xylem, a feature more commonly associated with cordaitaleans, such as *Mesoxylon* stem and the leaf traces of *Shanxioxylon* (Trivett & Rothwell, 1985; Wang *et al.*, 2025). At the same time, the general stem anatomy resembles the walchian *Thucydia* (see Hernández-Castillo et al., 2001; Shi-Jun et al., 2003). Stems and roots attributed to Cordaitales have already been reported from the region (Coimbra & Mussa, 1984; Conceição et al., 2020b), providing further support for the presence of this group. However, additional data are required to confirm the precise taxonomic placement of *Parnaiboxylon* within the coniferophytes.

Paulistoxylon was originally described from early Permian (Kungurian) deposits in southeastern Brazil (Mussa, 1986a). Subsequently, Crisafulli (1998) associated the genus with Cordaitales, based on its diaphragmed pith, as well as the dimensions and radial pitting patterns of the tracheids. All models previously trained in *PaleoWood* indicated the same taxonomic affinity consensually, further supporting this classification (Conceição et al., 2023). The SVM models, in contrast, were inconclusive, suggesting potential links to both Cordaitales or Voltziales (Tab. 3). Despite this ambiguity, the recurring association with Cordaitales across multiple analyses points to a prevailing affinity with this order. Nonetheless, the alternative prediction involving Voltziales may be of interest, as it could reflect shared anatomical traits or convergent features worthy of further investigation.

Teresinoxylon is a stem from the early Permian of northwestern Brazil, described as having a morphology intermediate between a mixed protostele and an eustele. It displays unorganized bundles of tracheids within the pith, both mesarch and endarch primary xylem, tall and wide parenchymatous rays, podocarpoid cross-field pits, and manoxylic wood (Caldas et al., 1989). According to Caldas et al. (1989), its anatomical features suggest affinities with *Cycadoxylon* and *Ptychoxylon*, which at the time were considered pteridosperms due to similarities with lyginopterids. *Teresinoxylon* indeed shares a set of characters comparable to several pteridosperms, such as *Diichnia* (Calamopityales) and *Heterangium* (Lyginopteridales) (see Read, 1936; Pigg et al., 1987). The SVM predictions support this hypothesis, suggesting an affinity with Lyginopteridales, a plausible taxonomic placement that is both morphologically consistent and reinforced by computational modeling (Tab. 3).

Overall, the application of the SVM classifier to Gondwanan fossil genera demonstrated its effectiveness in recovering well-supported taxonomic affinities, even in the absence of complete anatomical information. In cases such as *Amosioxylon*, *Iratinia*, and *Teresinoxylon*, the predictions were consistent with previously suggested affinities and supported by anatomical features recognized in the literature. For *Eoguptioxylon*, *Parnaiboxylon* and *Paulistoxylon*, although the results were not fully convergent, the suggested classifications remained within reasonable phylogenetic bounds, reflecting known anatomical complexities and gaps in current taxonomic resolution. These outcomes reinforce the potential of SVM as a valuable tool for assisting paleobotanical interpretations.

Model Improvements and Limitations

Compared to the original *PaleoWood* models, the SVM classifier demonstrated substantial improvements in both predictive accuracy and overall model performance (Tab. 2). Validation results indicate a better model fit and a more robust understanding of the relationship between anatomical

characters and taxonomic groups, even when working with limited and unbalanced fossil data. Moreover, the SVM exhibited stronger generalization capacity, accurately predicting complex cases and reflecting established anatomical groupings, as well as providing insights into the affinities of certain fossil genera (Tab. 3). Even so, challenges remain in the classification of fossil plants, which directly affect the performance of computational models that rely on such data.

The situation is even more complex for the Paleozoic flora of Gondwana, where whole-plant reconstructions are rare. Establishing precise taxonomic relationships among these organisms has proven difficult, often relying solely on morphological and anatomical comparisons with fossils of better-established phylogenetic placement. Clear affiliations between wood-derived taxa and their original mother plants are often missing (e.g., Guerra-Sommer, 1977; Mussa, 1978, 1986a, 1986b; Crisafulli, 1998, 2001; Merlotti, 2002, 2009; Kurzawe et al., 2013; Decombeix et al., 2019; Wan et al., 2020; Conceição et al., 2020a, 2022). In this context, machine-assisted analyses may offer valuable contributions by delivering classifications that are less prone to subjective bias. Although the methodological approach has shown considerable improvement, as evidenced by the SVM's performance compared to *PaleoWood*, the limited availability of well-resolved taxonomic data continues to constrain further advances in classification.

Entering more specifically into the field of applied artificial intelligence, recent reviews have demonstrated the advancement of its use in paleontology (see Yu et al., 2024; Liu et al., 2026). Among the proposed artificial intelligence models, expert systems were the first to be used in studies involving fossils. By definition, an expert system, as a case of a knowledge-based system, simulates expert decision-making by relying on structured domain knowledge and predefined rules. While useful for processing known inputs, they lack adaptability, struggle with unfamiliar data, and require manual updates as knowledge evolves. These limitations have reduced their use in favor of more flexible, data-driven approaches in artificial intelligence (Yu et al., 2024; Liu et al., 2026). Significant examples of expert systems include if-then structured models based on logical rules, which encode expert knowledge into decision pathways for automated reasoning (Russell & Norvig, 1995), such as computerized taxonomic tables and identification keys.

Notably, these reviews have compared *PaleoWood* to an expert system or a knowledge-based system (Yu et al., 2024; Liu et al., 2026), likely due to its limitations and unresolved issues. However, we argue that this comparison is not entirely appropriate, as both the data and algorithms employed herein and previously by Conceição et al. (2023) rely on machine learning and generalization, rather than on predefined rule-based logic. The selected input variables are relevant to plant taxonomy, though they are not necessarily ideal for distinguishing higher-level taxonomic groups, as they often reflect intrageneric variation. Although there are observable trends in character states across taxonomic orders and similarities in the anatomy, no single diagnostic trait used allows for a straightforward assignment to a specific group. There is no simple logic in the classification of axes and, complex comparisons are required, as done through SVM.

Another important issue is the limited dataset size, which results in a small number of training samples. As a consequence, the model may not be adequately fitted across the full vector space, requiring a high degree of generalization when confronted with atypical or previously unseen samples. This challenge is further compounded by the substantial anatomical variability within certain taxonomic orders, as well as by evolutionary convergences that produce character overlaps among phylogenetically distant groups. For instance, the voting mechanism occasionally adopted in multiclass problems, which includes SVM, is directly affected by the high similarity and internal variability of classes. When comparing pairs of classes in the one-against-one strategy, votes among anatomically similar groups may result in near ties, potentially leading to the selection of an incorrect category. Alternative strategies also face their own intricate limitations.

Taken together, these factors suggest that some inherent complexities of the dataset may not be fully resolved, regardless of algorithmic sophistication, although identifying these challenges guides the most appropriate choice of techniques. In such contexts, even advanced models like SVMs may reach a performance ceiling, not

necessarily due to computational limitations, but because of the biological and taphonomical constraints embedded in the data itself. Even so, such models can be highly valuable in detecting underlying patterns among plant groups and in supporting the delimitation of taxonomic and evolutionary hypotheses. By highlighting consistent character associations and revealing non-obvious affinities, computational approaches offer a complementary perspective to traditional paleobotanical analyses, which remain an irreplaceable basis in the study of fossil plants.

CONCLUSION

The SVM-based classifier successfully linked the anatomical features of Paleozoic gymnosperm axes to their respective taxonomical order. Compared to *Paleowood*, it showed improved overall performance, achieving a maximum accuracy of 71.4% and a maximum macro-averaged F1-Score of 61%, with regularization parameter C = 1 and C = 10, respectively. Cross-validation results suggest underfitting when $C \leq 0.1$, and overfitting if $C \geq 100$. Furthermore, the permutation test confirmed that the accuracy values reflect the model's learned structure rather than random associations. Together, these findings indicate that the SVM classifier offers a robust and reliable approach for extracting taxonomic information from anatomical data in fossil axes.

Building on these results, the classifier was applied to six fossil genera from western Gondwana, leading to insightful taxonomic predictions. For both *Amosioxylon* and *Teresinoxylon*, the model associated them with pteridosperms, a plausible hypothesis also discussed in their original descriptions. In contrast, *Eoguptioxylon*, *Parnaiboxylon*, and *Paulistoxylon* were linked to coniferophytes, although without a clear consensus on the specific taxonomic order to which they may belong. The predictions for *Parnaiboxylon* and *Paulistoxylon* align with prior literature, whereas the enigmatic *Eoguptioxylon* appears to diverge from currently known forms. Finally, *Iratinia*, previously suggested to be related to cycads, matches the SVM classification. These outcomes demonstrate the potential of the model to generate meaningful hypotheses and support paleobotanical interpretations, especially in cases where traditional classification remains uncertain.

The proposed classifier is grounded in machine learning and capable of generalizing beyond explicit logic, in contrast to expert systems. Even so, assigning affinities of fossil gymnosperm axes remains inherently difficult, as the variables used often capture variation at the genus or species level rather than providing clear distinctions among higher-level taxonomic groups. Even sophisticated methods like SVM face limitations when the available data are scarce or overlapping. Despite these obstacles, machine learning models still prove powerful in uncovering subtle patterns and supporting paleobotanical investigation. They do not replace researcher interpretation, but when well-applied, they can reveal latent patterns in the fossil record.

ACKNOWLEDGMENTS

We thank Alejandro Otero for the invitation to contribute to this thematic volume, and the invited editors and anonymous reviewers for their valuable contributions. This study was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001(process: DS 88887.699755/2022-00). DMC also thanks CAPES, Finance Code 001, and the Alexander von Humboldt Foundation for the funding support provided through the CAPES-Humboldt Research Fellowship Program, and RI thanks the National Council for Scientific and Technological Development for fellowships awarded (CNPq, process: PQ 313946/2021-3).

REFERENCES

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In J. F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Machine Learning: ECML 2004* (pp. 39–50). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30115-8_7

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In M. Awad, & R. Khanna (Eds.), *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 39–66). Apress. https://doi.org/10.1007/978-1-4302-5990-9_3

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. In H. He, & Y. Ma (Eds.) *Imbalanced learning: Foundations, Algorithms, and Applications*, (pp. 83–99). Wiley. https://doi.org/10.1002/9781118646106.ch5

Beck, C.B., Schmid, R., & Rothwell, G.W. (1982). Stelar morphology and the primary vascular system of seed plants. *Botanical Review, 48*(4), 691–815. https://doi.org/10.1007/BF02860874

Burbidge, R., & Buxton, B. (2001). An introduction to support vector machines for data mining. Keynote papers, young OR12, 3–15.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(*2*), 121–167. https://doi.org/10.1023/A:1009715923555

Caldas, E. B., Mussa, D., Lima Filho, F. P., & Rösler, O. (1989). Nota sobre a ocorrência de uma floresta petrificada de idade permiana em Teresina, Piauí. *Boletim IG-USP. Publicação Especial*, *7*, 69–87. https://doi.org/10.11606/issn.2317-8078.v0i7p69-87

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j. neucom.2019.10.118

Césari, S., Archangelsky, S., & Vega, J. (2005). Anatomy of a new probable pteridosperm stem from the Late Carboniferous of Argentina. *Revista del Museo Argentino de Ciencias Naturales*, *Nueva Serie*, *7*(*1*), 7–15.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(*3*), 1–27. https://doi.org/10.1145/1961189.1961199

Coimbra, A. M., & Mussa, D. (1984). Associação lignitafoflorística da Formação Pedra-de-Fogo, (Arenito Cacunda), Bacia do Maranhão-Piauí, Brasil. *Anais do XXXIII Congresso Brasileiro de Geologia* (v 2, pp. 591–605). Rio de Janeiro.

Conceição, D. M., Crisafulli, A., Iannuzzi, R., Neregato, R., Cisneros, J. C., & Andrade, L. S. (2020a). New petrified gymnosperms from the Permian of Maranhão (Pedra de Fogo formation), Brazil: *Novaiorquepitys* and *Yvyrapitys*. *Review of Palaeobotany and Palynology*, *276*, 104177. https://doi.org/10.1016/j.revpalbo.2020.104177

Conceição, D. M., Iannuzzi, R., Andrade, L. S., Esperança Júnior, M. G. F., Cisneros, J. C., & Siegloch, A. M. (2020b). First report of *Cordaixylon* Grand'Eury in the Permian of South America, Parnaíba Basin, Brazil. *Journal of South American Earth Sciences*, *101*, 102620. https://doi.org/10.1016/j.jsames.2020.102620

Conceição, D. M., Esperança Júnior, M. G. F., Iannuzzi, R., & Cisneros, J. C. (2022). Two new petrified gymnosperms with solenoid piths from the Pedra de Fogo Formation, Permian of Maranhão, Brazil. *Review of Palaeobotany and Palynology*, *299*, 104622. https://doi.org/10.1016/j.revpalbo.2022.104622

Conceição, D. M., Esperança Júnior, M. G. F., Iannuzzi, R., Recamonde-Mendoza, M., & Malta, G. B. B. O. (2023). *PaleoWood*: A machine learning approach for determining the affinity of Paleozoic gymnosperm woods. *Journal of South American Earth Sciences*, *121*, 104125. https://doi.org/10.1016/j.jsames.2022.104125

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Crisafulli, A. (1998). Leños gimnospérmicos de la Formación Melo (Pérmico Inferior), Uruguay. Parte III. *Paulistoxylon*, *Austrosclero-medulloxylon* y *Piracicaboxylon*. *Ameghiniana*, *35*(*2*), 217–225.

Crisafulli, A. (2001). Leños pérmicos de la Formación Yaguarí, República Oriental del Uruguay. (2014). *Ameghiniana*, *38*(*1*), 61–72.

Crisafulli, A., & Lutz, A. (2007). *Eoguptioxylon antiqua* gen. et sp. nov. (Pteridospermae) del Pérmico Superior (Formación la Antigua), provincia de la Rioja, Argentina. *Ameghiniana*, *44*(*1*), 197–204.

Cristianini, N., & Ricci, E. (2008). Support vector machines. In Kao, M.Y. (Ed.), *Encyclopedia of algorithms* (pp. 928–932). Springer. https://doi.org/10.1007/978-0-387-30162-4_415

Decombeix, A. L., Meyer-Berthaud, B., & Galtier, J. (2011). Transitional changes in arborescent lignophytes at the Devonian–Carboniferous boundary. *Journal of the Geological Society*, *168*(*2*), 547–557. https://doi.org/10.1144/0016-76492010-074

Decombeix, A. L., Galtier, J., McLoughlin, S., Meyer-Berthaud, B., Webb, G. E., & Blake, P. R. (2019). Early Carboniferous lignophyte tree diversity in Australia: Woods from the Drummond and Yarrol basins, Queensland. *Review of Palaeobotany and Palynology*, *263*, 47–64. https://doi.org/10.1016/j.revpalbo.2019.01.009

Dunn, M. T. (2006). A review of permineralized seed fern stems of the Upper Paleozoic. *The Journal of the Torrey Botanical Society*, *133*(*1*), 20–32. https://doi.org/10.3159/1095-5674(2006)133[20:AROPSF]2.0.CO;2

Ferguson, D. K. (2005). Plant taphonomy: ruminations on the past, the present, and the future. *Palaios*, *20*(*5*), 418–428. https://doi.org/10.2110/palo.2005.P05-25p

Galtier, J., & Meyer-Berthaud, B. (2006). The diversification of early arborescent seed ferns. *The Journal of the Torrey Botanical Society*, *133*(*1*), 7–19. https://doi.org/10.3159/1095-5674(2006)133[7:TDOEAS]2.0.CO;2

Greenwood, D. R. (1991). The taphonomy of plant macrofossils. In S.K. Donovan (Ed.), *The Processes of Fossilization* (pp. 141–169), Columbia University Press.

Guerra-Sommer, M. (1977). *Damudoxylon* (Maheshwari) Maheshwari, 1972, um gênero ocorrente no Gondwana do Brasil. *Pesquisas em Geociências*, *7*(*7*), 131–144. https://doi.org/10.22456/1807-9806.21824

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(*4*), 18–28. https://doi.org/10.1109/5254.708428

Hernández-Castillo, G. R., Rothwell, G. W., & Mapes, G. (2001). Thucydiaceae fam. nov., with a review and reevaluation of Paleozoic walchian conifers. *International Journal of Plant Sciences*, *162*(*5*), 1155–1185. https://doi.org/10.1086/321920

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A practical guide to support vector classification*. Department of Computer Science and Information Engineering, National Taiwan University, Taipei.

Iannuzzi, R., & Vieira, C. E. L. (2005). *Paleobotânica*. Editora da UFRGS.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & Rajabi, M.J. (2014). Advantage and drawback of support vector machine functionality. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 63–65. https://doi.org/10.1109/I4CT.2014.6914146

Kurzawe, F., Iannuzzi, R., Merlotti, S., Rößler, R., & Noll, R. (2013). New gymnospermous woods from the Permian of the Parnaíba Basin, northeastern Brazil, part I: *Ductoabietoxylon*, *Scleroabietoxylon* and *Parnaiboxylon*. *Review of Palaeobotany and Palynology*, *195*, 37–49. https://doi.org/10.1016/j.revpalbo.2012.12.004

Lacey, W. S. (1953). Scottish Lower Carboniferous Plants: *Eristophyton waltoni* sp. nov. and *Endoxylon zonatum* (Kidston) Scott from Dunbartonshire. *Annals of Botany*, *17*(*4*), 579–597. https://doi.org/10.1093/oxfordjournals.aob.a083371

Liu, Y., Torres, L. N., Wang, B., Pei, W., Na, Y., Song, Q., & Shi, X. (2026). Artificial Intelligence in Paleobotany and Palynology. *Geological Journal*. *61*(3), 748–769. https://doi.org/10.1002/gj.70007

Lorena, A. C., & De Carvalho, A. C. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, *14*(*2*), 43–67. https://doi.org/10.22456/2175-2745.5690

Ma, Y., & Guo, G. (2014). *Support vector machines applications*. Springer. https://doi.org/10.1007/978-3-319-02300-7

MacFadden, B. J., Barberis, C. A., Vallejo-Pareja, M. C., Zbinden, S. P., Perez, V. J., Killingsworth, S. R., Marks, K. W., Hall, D., & Porto, A. (2026). AI and paleontology: effects of vertebrate fossil sample size on machine learning image classification. *Paleobiology*, 1–17. https://doi.org/10.1017/pab.2025.10084

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(*3*), 283–289. https://doi.org/10.1002/wics.49

Martín-Perea, D. M., Courtenay, L. A., Domingo, M. S., & Morales, J.

(2020). Application of artificially intelligent systems for the identification of discrete fossiliferous levels. *PeerJ*, *8*, e8767. https://doi.org/10.7717/peerj.8767

Merlotti, S. (2002). Dois novos taxa lenhosos da Formação Serra Alta (Permiano Superior, Bacia do Paraná), estado do Rio Grande do Sul, Brasil. *Revista Geociências – UNG-Ser – ISSN 1981-741X (1981-7428)*, *7*(*6*), 5–14.

Merlotti, S. (2009). Reavaliação taxonômica de lenhos das formações Irati e Serra Alta, Permiano da Bacia do Paraná, Brasil. *Pesquisas em Geociências*, *36*(*1*), 11–21. https://doi.org/10.22456/1807-9806.17871

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2024). *Package 'e1071'*. https://doi.org/10.32614/CRAN.package.e1071

Moguerza, J. M., & Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, *21*(*3*), 322–336. https://doi.org/10.1214/088342306000000493

Mussa, D. (1978). *Brasilestiloxylon* e *Solenobrasilioxylon*, dois novos gêneros gondwânicos na formação Irati, *Boletim IG-USP*, *9*, 118–127. https://doi.org/10.11606/issn.2316-8978.v9i0p118-127

Mussa, D. (1986a). Eustelos Gondwânicos de medulas diafragmadas e a sua posição estratigráfica. *Boletim IG-USP. Série Científica*, *17*, 11–24. https://doi.org/10.11606/issn.2316-8986.v17i0p11-24

Mussa, D. (1986b). As formas gondwânicas do Grupo Solenóide e sua distribuição estratigráfica. *Anais da Academia Brasileira de Ciências*, *58*(*1*), 61–88.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of machine learning research*, *11*(62), 1833–1863.

Philippe, M., Afonin, M., Delzon, S., Jordan, G. J., Terada, K., & Thiébaut, M., (2019). A paleobiogeographical scenario for the Taxaceae based on a revised fossil wood record and embolism resistance. *Review of Palaeobotany and Palynology*, *263*, 147–158.

Pigg, K. B., Taylor, T. N., & Stockey, R. A. (1987). Paleozoic seed ferns: *Heterangium kentuckyensis* sp. nov., from the Upper Carboniferous of North America. *American Journal of Botany*, *74*(*8*), 1184–1204. https://doi.org/10.1002/j.1537-2197.1987.tb08733.x

Platt, J. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Microsoft Research.

Posit Team (2024). *RStudio: Integrated Development Environment for R*. Posit Software, PBC.

Read, C. B. (1936). The flora of the New Albany shale: Part 1, *Diichnia kentuckiensis*, a new representative of the Calamopityeae. *U.S. Geological Survey professional paper*, *185–H*, 149–161.

Rößler, R., Philippe, M., van Konijnenburg-van Cittert, J. H., McLoughlin, S., Sakala, J., Zijlstra, G., Annamraju, R., Ash, S., Baas, P., Bamford, M., Bateman, R., Booi, M., Boonchai, N., Brea, M., Crisafulli, A., Decombeix, A. L., Dolezych, M., Dutra, T., Esteban, L.G., ... Zhang, W. (2014). Which name (s) should be used for *Araucaria*-like fossil wood?—Results of a poll. *Taxon*, *63*(1), 177–184. https://doi.org/10.12705/631.7

Russell, S.J., & Norvig, P. (1995). *Artificial Intelligence. A Modern Approach. Second Edition*. Prentice Hall.

Shi-Jun, W., Hilton, J., Baolin, T., & Galtier, J. (2003). Cordaitalean seed plants from the Early Permian of North China. I. Delimitation and reconstruction of the *Shanxioxylon sinense* plant. *International Journal of Plant Sciences*, *164*(*1*), 89–112. https://doi.org/10.1086/344558

Silvestro, D., & Pimiento, C. (2025). Emerging uses of artificial intelligence in deep time biodiversity research. *Nature Reviews Biodiversity. 1*, 671–677. https://doi.org/10.1038/s44358-025-00075-4

Simões, M. G., Rodrigues, S. C., Bertoni-Machado, M., & Holz, M. (2010). Tafonomia: processos e ambientes de fossilização. In I. S. Carvalho (Ed.), *Paleontologia* (pp. 19–51), Interciência.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Spicer, R. A. (1991). Plant taphonomic processes. In P. A. Allison & D. Briggs (Eds.), *Taphonomy: Releasing the data locked in the fossil record* (pp. 71–113), Springer.

Spiekermann, R., Jasper, A., Siegloch, A. M., Guerra-Sommer, M., & Uhl, D. (2021). Not a lycopsid but a cycad-like plant: *Iratinia australis* gen. nov. et sp. nov. from the Irati Formation, Kungurian of the Paraná Basin, Brazil. *Review of palaeobotany and palynology*, 289, 104415. https://doi.org/10.1016/j.revpalbo.2021.104415

Taylor, E. L., Taylor, T. N., & Krings, M. (2009). *Paleobotany: the biology and evolution of fossil plants*. Academic Press.

Trivett, M. L., & Rothwell, G. W. (1985). Morphology, systematics, and paleoecology of Paleozoic fossil plants: *Mesoxylon priapi*, sp. nov.(Cordaitales). *Systematic Botany*, *10*, 205–223. https://doi.org/10.2307/2418347

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, *6*, 1453–1484.

Wan, M., Shi, G. R., Luo, M., Lee, S., & Wang, J. (2020). First record of a petrified gymnospermous wood from the Kungurian (late Early Permian) of the southern Sydney Basin, southeastern Australia, and its paleoclimatic implications. *Review of Palaeobotany and Palynology*, *276*, 104202. https://doi.org/10.1016/j.revpalbo.2020.104202

Wang, K., Huang, X., Yang, W., Wang, J., & Wan, M. (2023). A new gymnospermous stem from the Moscovian (Carboniferous) of North China, and its palaeoecological significance for the Cathaysian Flora at the early evolutionary stage. *Review of Palaeobotany and Palynology*, *311*, 104858. https://doi.org/10.1016/j.revpalbo.2023.104858

Wang, K., Jia, G., Dong, L., Wang, J., Wang, S., Wang, J., & Wan, M. (2025). *Shanxioxylon yangquanense* sp. nov., a new Kasimovian cordaitalean axis from the Benxi Formation (Pennsylvanian, Carboniferous) of Yangquan City, Shanxi Province, North China. *Review of Palaeobotany and Palynology*, *335*, 105287. https://doi.org/10.1016/j.revpalbo.2025.105287

Yaqoob, M., Ishaq, M., Ansari, M. Y., Qaiser, Y., Hussain, R., Rabbani, H. S., Garwood, R. J., & Seers, T. D. (2025). Advancing paleontology: a survey on deep learning methodologies in fossil image analysis. *Artificial Intelligence Review*, *58*, 83. https://doi.org/10.1007/s10462-024-11080-y

Yu, C., Qin, F., Watanabe, A., Yao, W., Li, Y., Qin, Z., Liu, Y., Wang, H., Jiangzuo, Q., Hsiang, A. Y., Ma, C., Rayfield, E., Benton, M.J., & Xu, X. (2024). Artificial intelligence in paleontology. *Earth-Science Reviews*, 252, 104765. https://doi.org/10.1016/j.earscirev.2024.104765

doi: 10.5710/PEAPA.25.06.2026.562

Recibido: 9 de enero 2026
Aceptado: 25 de junio 2026
Publicado: 9 de septiembre 2026

APPENDIX 1

Testing more complex kernels using default parameter settings of e1071 package in R. This was implemented by replacing "linear" with "polynomial", "radial" and "sigmoid" in L. 40 of the script (https://github.com/marioesperancajr/SVMwood).

	C (polynomial)						
	1	10	50	75	85	100	150
Accuracy	0.14	0.38	0.69	0.69	0.71	0.71	0.67
	C (radial basis)						
	0.1	1	5	10	30	50	100
Accuracy	0.21	0.36	0.50	0.69	0.64	0.64	0.64
	C (sigmoid)						
	1	5	10	50	100	500	1000
Accuracy	0.33	0.48	0.45	0.69	0.67	0.64	0.67