BIULETYN INSTYTUTU HODOWLI I AKLIMATYZACJI ROŚLIN Nr 303/304 / 2025 : 19–33
E-ISSN: 2657–8913 DOI: 10.37317/biul-2025-0006

Oryginalny Artykuł Naukowy
Original Research Paper

Statistical analysis of seeds morphology and texture for interspecific similarity assessment across taxonomic levels

Analiza statystyczna morfologii i tekstury nasion w celu oceny podobieństwa międzygatunkowego na różnych poziomach taksonomicznych

Seweryn Lipiński ✉

Faculty of Technical Sciences, University of Warmia and Mazury in Olsztyn

✉seweryn.lipinski@uwm.edu.pl

Seed-based identification is important for both botanical research and agricultural practice, yet many recent approaches rely on opaque deep learning models. This study evaluates whether simple geometric and textural descriptors extracted from seed images can capture taxonomic patterns across species, genera and families. A dataset of 4,496 seed images representing 88 plant species was analysed using 13 interpretable shape- and texture-based features. Multivariate Analysis of Variance (MANOVA), canonical variate analysis and Mahalanobis distances were applied to quantify group separation, while hierarchical clustering and heatmaps served as complementary exploratory tools to visualize similarity structures. Across all taxonomic levels, MANOVA revealed highly significant multivariate differences, with most separation explained by a consistent subset of shape-related descriptors, particularly solidity, extent, convex area and major axis length. Texture features contributed only marginally. These findings indicate that fundamental, explainable seed morphology descriptors provide a robust and scalable basis for taxonomic differentiation, offering an interpretable alternative to black-box classification approaches and supporting automated plant identification.

Keywords: seed image analysis, plant species identification, botanical classification, geometric and textural features, hierarchical clustering

Identyfikacja na podstawie cech nasion jest ważna zarówno dla badań botanicznych, jak i praktyki rolniczej, jednak wiele współczesnych podejść opiera się na nieprzejrzystych modelach głębokiego uczenia. Niniejsze badanie ocenia, czy proste deskryptory geometryczne i teksturalne wyodrębnione z obrazów nasion mogą uchwycić wzorce taksonomiczne w obrębie gatunków, rodzajów i rodzin. Zbiór danych 4496 obrazów nasion reprezentujących 88 gatunków roślin został przeanalizowany z wykorzystaniem 13 interpretowalnych cech opartych na kształcie i teksturze. Wielowymiarowa analiza wariancji (MANOVA), analiza zmiennych kanonicznych oraz odległości Mahalanobisa zostały zastosowane do ilościowego określenia separacji grup, podczas gdy hierarchiczna analiza skupień i mapy cieplne (heatmaps) posłużyły jako uzupełniające narzędzia eksploracyjne do wizualizacji struktur podobieństwa. Na wszystkich poziomach taksonomicznych MANOVA ujawniła wysoce istotne różnice wielowymiarowe, przy czym większość separacji została wyjaśniona przez spójny podzbiór deskryptorów związanych z kształtem, w szczególności zwartość, powierzchnię, powierzchnię wypukłą i długość osi głównej. Cechy tekstury miały jedynie marginalny wpływ. Wyniki te wskazują, że podstawowe deskryptory morfologii nasion zapewniają solidną i skalowalną podstawę do różnicowania taksonomicznego, oferując interpretowalną alternatywę dla podejść klasyfikacyjnych typu „czarna skrzynka” i wspierając automatyczną identyfikację roślin.

Słowa kluczowe: analiza obrazu nasion, identyfikacja gatunków roślin, klasyfikacja botaniczna, cechy geometryczne i teksturalne, hierarchiczna analiza skupień

Introduction

In recent years, the field of seed recognition and classification has closely followed the broader trends observed in image analysis and pattern recognition. Deep learning approaches, particularly convolutional neural networks (CNNs), have become the dominant methodology (Chen et al., 2021; Taye, 2021, Zhao et al., 2024). These networks have been applied successfully to various seed-related tasks, such as classifying seeds of different species (Eryigit and Tugrul, 2021; Kumar et al., 2024; Loddo et al., 2021) and differentiating between varieties within a single species. Examples include maize variety recognition (Wang and Wang, 2021), wheat cultivar classification (Yasar, 2024), rice seed analysis (Rajalakshmi, 2024), chickpea sorting (Taheri-Garavand et al., 2021), and the identification of cannabis seeds (Islam et al., 2024).

Despite the remarkable classification performance of neural networks, especially in large-scale applications, their use in scientific studies is not without drawbacks. One fundamental limitation is their dependency on the size, quality, and representativeness of the training datasets. In botanical contexts, where image data might be scarce or imbalanced across species, this can limit their generalizability. Another critical issue is the lack of interpretability – CNNs are often regarded as

Redaktor Wiodący / Leading Editor
Dariusz R. Mańkowski

"black-box" models, providing little or no insight into the specific features used for decision-making (Buhrmester et al., 2021, Szandała, 2023). As neural networks become deeper and more complex, they tend to rely on increasingly abstract representations that are difficult to relate to biological or morphological traits (Barbierato and Gatti, 2024; Krohn et al., 2019). This poses challenges for biological research, where transparency, reproducibility, and biological plausibility are essential.

Alongside deep learning, more interpretable approaches to seed image analysis remain active in literature. For example, methods involving handcrafted features have been successfully used to classify pepper seeds using colour filter array images (Djoulde, 2024), assess grape seed morphology (Espinosa-Roldán et al, 2024), analyse pumpkin seed diversity (Ermiş et al., 2025), and detect broken soybean seeds (Chen et al., 2022). These approaches offer the advantage of using well-defined, explainable features, such as shape descriptors, colour statistics, and textural measures, which can be directly visualized, measured, and interpreted. They also allow us to identify and communicate which features are most relevant to classification decisions, supporting scientific understanding and practical implementation.

However, such methods often rely on highly customized descriptors tailored to specific tasks, limiting their general applicability. This raises the question of whether a consistent set of fundamental geometrical and textural features, extracted from binary or grayscale images of seeds, might be sufficient to differentiate species across taxonomic levels. If so, this would enable a transparent and accessible approach to seed classification, applicable in various biological and ecological contexts, without the need for extensive computational resources or large annotated datasets.

Motivated by this hypothesis, the present study explores the effectiveness of basic morphological and texture-based features in distinguishing seeds of different plant taxa. The proposed approach involves extracting a set of general-purpose descriptors from seed images and subjecting them to statistical analysis in order to quantify interspecific, intergeneric, and interfamilial similarities and differences. This enables both numerical comparisons and visual interpretations of patterns in seed morphology and texture.

The key aim of this article is to evaluate whether simple, quantifiable features can reveal meaningful biological patterns across species, genera, and families, and whether these patterns are consistent with established taxonomic relationships. By focusing on explainable descriptors and standard image analysis techniques, the study contributes a methodology that complements existing black-box approaches, while offering valuable insights into seed morphology and its potential for species identification and classification.

Materials and Methods

Seed dataset

The seed images analysed in this study originate from a publicly available dataset (A dataset based on smartphone acquisition that can be used for seed identification using deep learning models, 2023), with a description presented in (Yuan et al., 2024).

The full dataset comprises 4,496 images representing seeds from 88 plant species. Each image has a fixed resolution of 192×272 pixels and depicts a single seed photographed under standardized conditions. Although the acquisition process was standardized, natural differences in seed size, surface properties, and lighting resulted in varying image quality. This variability was intentionally preserved, as it reflects realistic conditions that may occur in practical applications and ensures that the proposed approach is robust to such differences. Fig. 1 presents a selection of eight representative seed images from this dataset, along with their corresponding species names.

As previously mentioned, the dataset includes seeds from 88 plant species, with a strong representation from the legume (Leguminosae) and grass (Poaceae) families - these two families account for exactly half of the species in the dataset.

It should be noted that the actual content of the dataset does not fully match its description in the original article. This discrepancy explains the slight differences between the data presented in Fig. 2 and the information reported in the paper describing the dataset (Yuan et al., 2024). However, the dataset used in this study is the original version provided by the authors, without any modifications.

Used software

MATLAB software was utilized for both image and statistical analysis, as well as for results visualization, with particular emphasis on the Statistics and Image Processing Toolboxes (Cho and Martinez, 2014; Gonzalez et al., 2003; Reyes-Aldasoro, 2025). Fig. 3 presents the workflow of the experimental design and data analysis.

2.3. Geometric and textural characteristics

A variety of properties were determined for the seed images. For texture analysis, mean, standard deviation, and entropy were calculated from grayscale versions of the images, as these measures are based on the normalized gray-level histogram. These features were selected because they capture fundamental aspects of seed surface texture: mean reflects overall brightness, standard deviation indicates contrast and variability, and

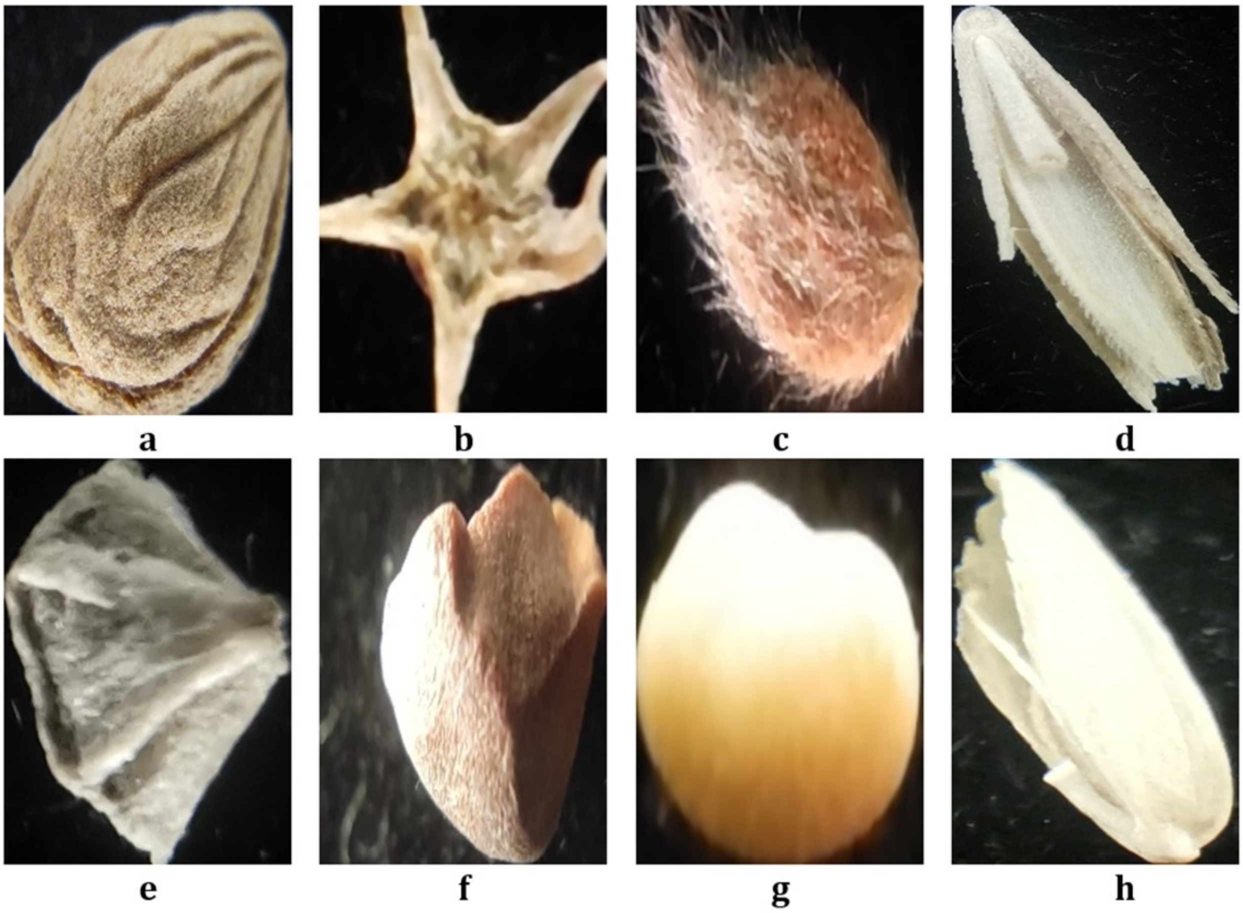

Fig. 1. Exemplary seed images from the used dataset: *Amygdalus mongolica* (a), *Bassia dasyphylla* (b), *Clematis fruticose* (c), *Elymus sibiricus* (d), *Halostachys caspica* (e), *Iris lactea* (f), *Medicago sativa* (g), and *Poa annua* (h).

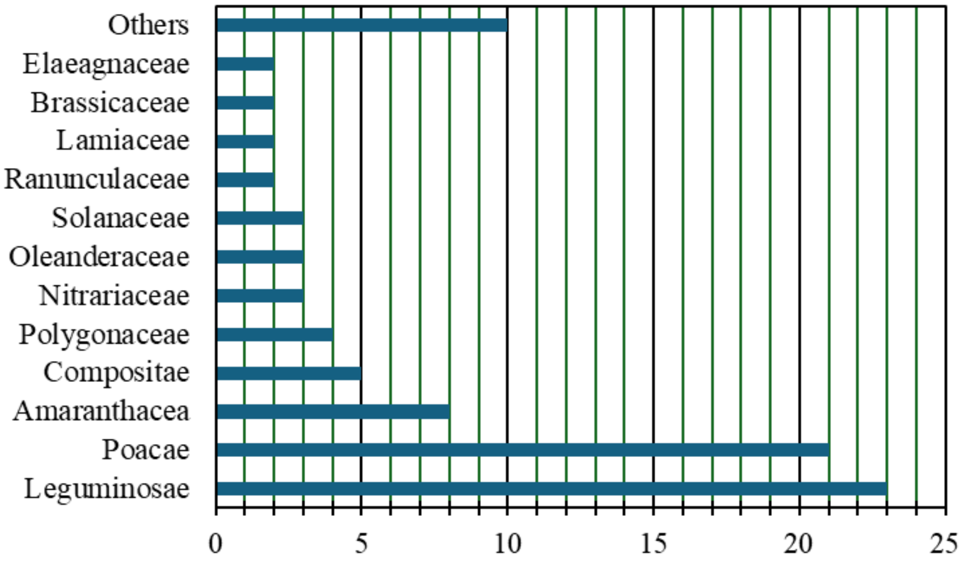

Fig. 2. Number of species in each plant family represented in the dataset used in this study.

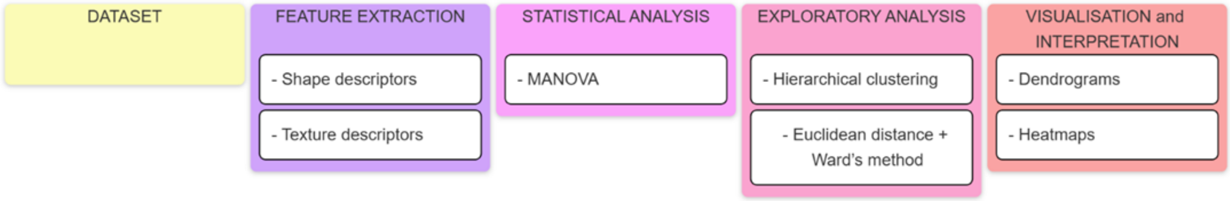

Fig. 3. Workflow of the analysis steps.

entropy measures complexity and irregularity of the surface.

For shape analysis, the following properties were calculated from binary images: Area, Euler number (equal to the number of holes inside a binarized seed image), Extent (the ratio of the area of the object to the area of its axis-aligned bounding box), Perimeter (the number of pixels in the boundary between a seed and the background), Convex area (the area of the convex hull of the seed; the convex hull is the smallest convex shape that can completely enclose the seed), Filled area (the number of pixels in the seed after filling any holes within its binarized image), Solidity (the proportion of the pixels in the convex hull that are also in the region), Eccentricity (calculated from the fitted ellipse derived from image moments; defined as the ratio of the distance between the foci of the ellipse and its major axis length, which reflects how elongated the seed shape is), Equivalent diameter (the diameter of a circle with the same area as the seed), and Major and Minor axis length (the lengths of the fitted ellipse axes based on image moments) (Gonzalez et al., 2003; Reyes-Aldasoro, 2025).It should be noted that the above parameters are independent of seed orientation. Therefore, the Major and Minor axis length parameters refer directly to the longer and shorter axes of the seed, rather than the x and y axes.

The binarization of colour images to calculate properties of binary images was performed using the Otsu automatic thresholding method (Otsu, 1979; Sezgin and Sankur, 2024). This method is one of the most frequently used and cited thresholding techniques (Kalicka and Lipiński, 2010; Lipiński and Lipiński, 2020). Its other advantages include ease of implementation and low computational demand.

After calculating the set of features for each seed, matrices were created where each row contained the features of a specific seed, labelled with its species. This formed the input data for statistical analysis.

Statistical analysis and results visualization

The experimental workflow combined confirmatory and exploratory approaches to address both statistical and biological questions. First, Multivariate Analysis of Variance (MANOVA) was applied to test the hypothesis that predefined groups (species and families) differ significantly in multivariate space based on all extracted descriptors. The null hypothesis assumed no differences among groups. MANOVA is an extension of the ANOVA test that assesses the impact of independent variables on multiple dependent variables simultaneously. This method is particularly useful in studies involving several related outcome variables that may be influenced by the same factors (e.g., seed characteristics) (Wiesnerova and Wiesner, 2008). Wilks' Lambda was used as the test statistic to assess group differences.

Following the analysis, hierarchical clustering and heatmaps were employed as complementary exploratory tools to visualize similarity patterns among species and families. Hierarchical clustering was performed using Euclidean distance as the dissimilarity measure and Ward's linkage method for cluster formation. A dendrogram is a tree-like diagram that illustrates the arrangement of clusters produced by hierarchical clustering. It represents relationships among species based on Euclidean distance and Ward's linkage method, grouping species according to similarity in descriptor space (Pavlopoulos et al., 2010). Each branch of the dendrogram corresponds to a cluster, and the length of the branches indicates the dissimilarity between clusters, which in this context reflects the degree of inter-species variation in the calculated feature vectors.

Heatmaps provide a visual representation of data patterns and relationships within a dataset. Each cell in the heatmap corresponds to a specific feature value, with colours indicating its magnitude (Engle et al., 2017; Wilkinson and Friendly, 2009). When combined with clustering techniques, heatmaps can highlight groups of species with similar characteristics. Rows and columns were reordered based on clustering results, making it easier to identify patterns and associations among features. Heatmaps simplify the interpretation of complex, high-dimensional data by providing an intuitive and immediate visual summary. In this study, heatmaps were used to analyse the distribution of seed descriptors across species and families, supporting the identification of biologically meaningful patterns such as trait similarities within families.

All feature values visualized in the heatmaps were z-score standardized prior to clustering and visualization. Consequently, the colour scale reflects standardized values rather than raw measurements, ensuring comparability across variables with different scales. The intensity of colours indicates the relative magnitude of each standardized feature value for a given object; however, it does not represent the contribution or weight of individual descriptors to the clustering process. Hierarchical clustering relies on the joint multivariate structure of all descriptors, whereas the heatmap serves exclusively as a visualization of their distribution. Proximity in the dendrogram reflects statistical association among descriptors rather than direct biological or causal relationships.

It is also important to note that neither hierarchical clustering nor heatmaps constitute a method for feature selection. Assessing the relevance, redundancy, or contribution of variables requires dedicated analytical approaches, such as e.g. Principal Component Analysis. Furthermore, because many morphological descriptors may exhibit collinearity, removing features identified as highly similar should be approached with caution, as this may inadvertently suppress complementary information or distort the multivariate relationships. While some features may be correlated, their removal requires caution, as it could distort the overall representation of seed morphology, particularly for variables that may be more sensitive to variability in image quality.

Selection Criteria

For the analyses conducted at the genus and family levels, all available seed images belonging to a given genus or family were included without any additional labels or filtering. This approach ensured that the clustering was based solely on the visual characteristics of the seeds, without introducing external classification constraints.

In contrast, the species-level analysis was intended as a preliminary test to demonstrate the applicability of the proposed tool and to verify whether species belonging to the same family exhibit greater similarity to each other than to species from different families. For this purpose, a small set of species was selected in alphabetical order, without applying any biological or ecological criteria. This simplified selection allowed for an initial assessment of the clustering performance and the potential of the method for more detailed taxonomic studies.

Results and Discussion

The following subsections analyse exemplary and representative clustering results, considering species (3.1), genera (3.2), and families (3.3). This analysis demonstrates the potential for examining differences in seeds features at various levels.

Two example analyses of differences and similarities at species level

First example

This subsection presents interspecific seed differentiation using six representative species as examples. To maintain objectivity, the first six species from the collection were selected in alphabetical order. These include *Achnatherum inebrians*, *Achnatherum splendens*, *Agropyron cristatum*, *Agropyron elongatum*, *Agropyron mongolicum* (all belonging to the Poaceae family), and *Agriophyllum squarrosum* (from the Amaranthaceae family).

Fig. 4 shows a dendrogram generated based on the statistical analysis of seed feature vectors. For visual reference, this figure also includes representative seed images corresponding to each species in the dendrogram. As previously mentioned, the y-axis of a dendrogram reflects the degree of dissimilarity between clusters - the higher the value, the greater the dissimilarity.

According to the dendrogram, *Agropyron mongolicum* and *Agropyron elongatum* form a tight cluster, indicating a high degree of similarity in seed features. While *Agropyron cristatum* belongs to the same genus, it appears more distinct and forms its own cluster, suggesting greater differentiation. These relationships are visually supported by the accompanying seed images, which confirm the level of similarity described by the dendrogram (it is worth recalling here that the calculated parameters are independent of seeds orientation in the photos). This proves the effectiveness of the feature vector in accurately capturing both similarities and differences among the seeds.

The *Agropyron* cluster shows some similarity to the *Achnatherum* group but remains clearly separate. Notably, although *Achnatherum splendens* and *Achnatherum inebrians* form a single cluster, *Achnatherum splendens* appears to have an equal level of similarity to both *Achnatherum inebrians* and the *Agropyron* cluster. This can be explained by the seed images: while seeds from the *Achnatherum* genus share similar surface textures, the seed of *Achnatherum inebrians* is visibly less slender than that of *Achnatherum splendens*, more closely resembling the compact shape of *Agriophyllum squarrosum*.

Agriophyllum squarrosum is clearly the most distinct among the analysed species, forming a separate branch in the dendrogram, which suggests unique morphological traits. This distinction is easily justified by its taxonomic classification - it belongs to the Amaranthaceae family, whereas all other analysed species are members of the Poaceae family.

The multivariate analysis of variance (MANOVA) confirmed statistically significant differences among the studied species ($p < 0.001$). The canonical analysis revealed that the first three canonical variates, associated with eigenvalues 4.04, 1.42, and 0.46, accounted for the vast majority of multivariate separation, whereas the remaining variates contributed negligibly. Pairwise Mahalanobis distances between group centroids further demonstrated clear interspecific structure. The closest species pair was *Agropyron elongatum* and

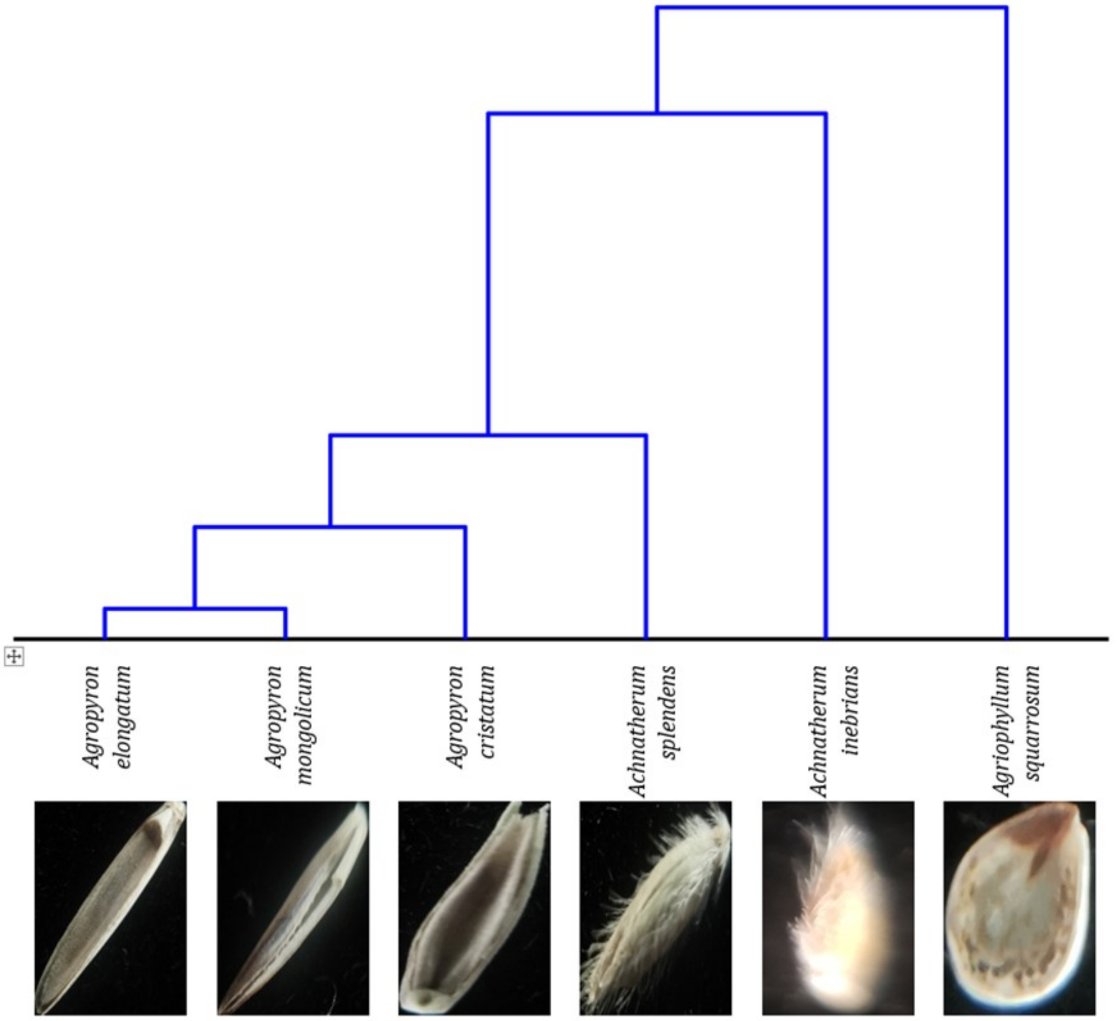

Fig. 4. Dendrogram obtained for six exemplary seeds species, and representative photos of seeds from the six chosen species.

Agropyron mongolicum (d = 2.97), followed by *A. cristatum–A. elongatum* (d = 4.58) and *A. cristatum–A. mongolicum* (d = 4.77), indicating a compact morphological cluster within the genus *Agropyron*. In contrast, *Agriophyllum squarrosum* was the most divergent taxon, exhibiting the largest distances to all remaining groups, particularly to *Achnatherum inebrians* (d = 30.24) and A. *splendens* (d = 28.28). These results demonstrate that most interspecific variation is captured by a small number of canonical dimensions and that the genus *Agropyron* forms a tightly cohesive multivariate group, while *A. squarrosum* remains clearly isolated.

The canonical variates were strongly associated with morphological descriptors related to seed shape and outline. CV1 was primarily driven by geometry–related features, including Major Axis Length, Solidity, Convex Area, Extent and Perimeter, indicating that overall shape proportions represented the dominant axis of interspecific separation. CV2 captured additional differences in convexity and elongation patterns, while CV3 was almost entirely determined by Major Axis Length, Solidity and Perimeter. In contrast, texture-based descriptors such as entropy and grey-level statistics had negligible loadings across all canonical variates. These results show that morphological rather than textural traits are the principal sources of multivariate differentiation among the studied species.

Fig. 5 shows a heatmap, being the result of analysis of individual seeds features. It allows us to observe the clusters formed by the seed features, which, for example, could be useful when creating a classification algorithm - it can help us identify features that are very similar or even redundant. The colour scale (in this heatmap, as well as in every subsequent one) ranges from -3 to 3, with red indicating higher values, green indicating lower values, and black representing intermediate values.

Analysis of the heatmap reveals that the Area and Filled Area parameters are closely related, which aligns with intuitive expectations. In principle, this suggests that only one of these features may be sufficient for classification. Less intuitively, a similar proximity is observed between Convex Area and Minor Axis Length.

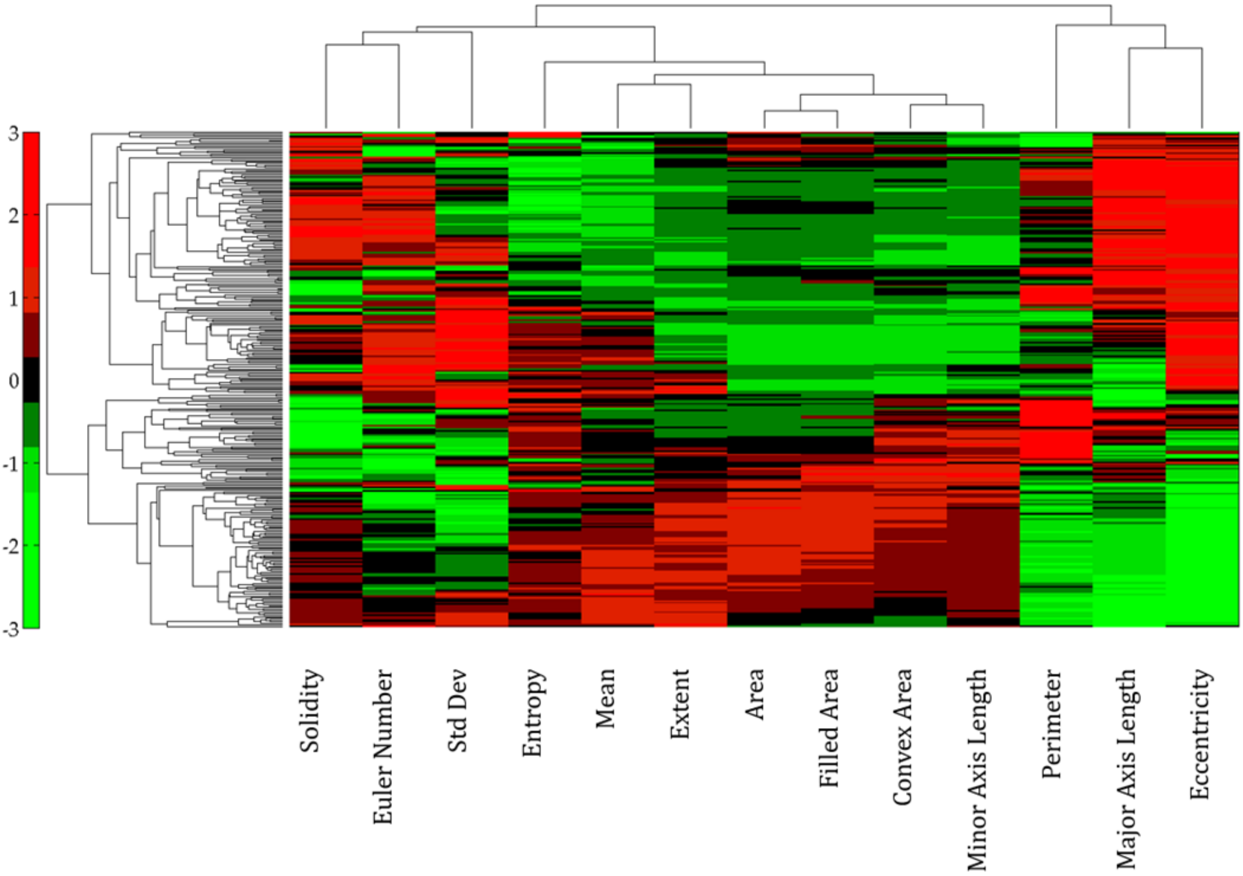

Fig. 5. Hierarchical clustering heatmap of seed features for six exemplary seeds species.

However, this proximity reflects similarity in the standardized feature values rather than their predictive relevance. Heatmaps and hierarchical clustering visualize similarity structures but do not evaluate feature importance.

For this reason, decisions regarding feature retention should be based on the performance of supervised classification models rather than on the exploratory visualizations. In these models, features such as Solidity, Euler Number, Standard Deviation, Perimeter, Major Axis Length, and Eccentricity contributed measurably to predictive accuracy and therefore cannot be removed without compromising performance.

Nonetheless, it is important to note that in specific cases - where different species share similar values for most features - subtle variations in brightness may render the distinction between Area and Filled Area critical for accurate classification. Therefore, feature reduction should be approached cautiously. This is particularly relevant considering that binary image-based features are computationally inexpensive, meaning that the potential benefits of reducing the feature count may not outweigh the risk of degrading classification accuracy.

Second example

To evaluate whether the proposed approach enables similarly intuitive interpretation for species from different families, an additional analysis was conducted using six species selected from the end part of the dataset.

The analyzed species included *Trifolium repens*, *Vicia sativa*, and *Vicia villosa* (all from the Leguminosae family), *Saposhnikovia divaricata* (Apiaceae), *Triticale* (Poaceae), and *Zygophyllum xanthoxylon* (Zygophyllaceae). Notably, this group represents greater taxonomic diversity than the previous set, making it a suitable basis for comparison.

Fig. 6 presents the dendrogram generated based on statistical analysis of seed morphological and textural feature vectors for this set of species. As in the case of Fig. 4, for visual reference, this figure includes representative seed images corresponding to each species in the dendrogram as well.

The dendrogram shows the hierarchical similarity structure among the analysed species based on their standardized seed descriptors. The first cluster to form is the pair *Triticale* and *Zygophyllum xanthoxylon*, which merge at the lowest dissimilarity level, indicating that these two species share the most similar descriptor profiles within the dataset. This cluster is subsequently joined by *Saposhnikovia divaricata*, forming a distinct group separate from the remaining species.

On the opposite branch, containing, as it turned out, members of the Leguminosae family, *Vicia sativa* and *Vicia villosa* form the closest pair,

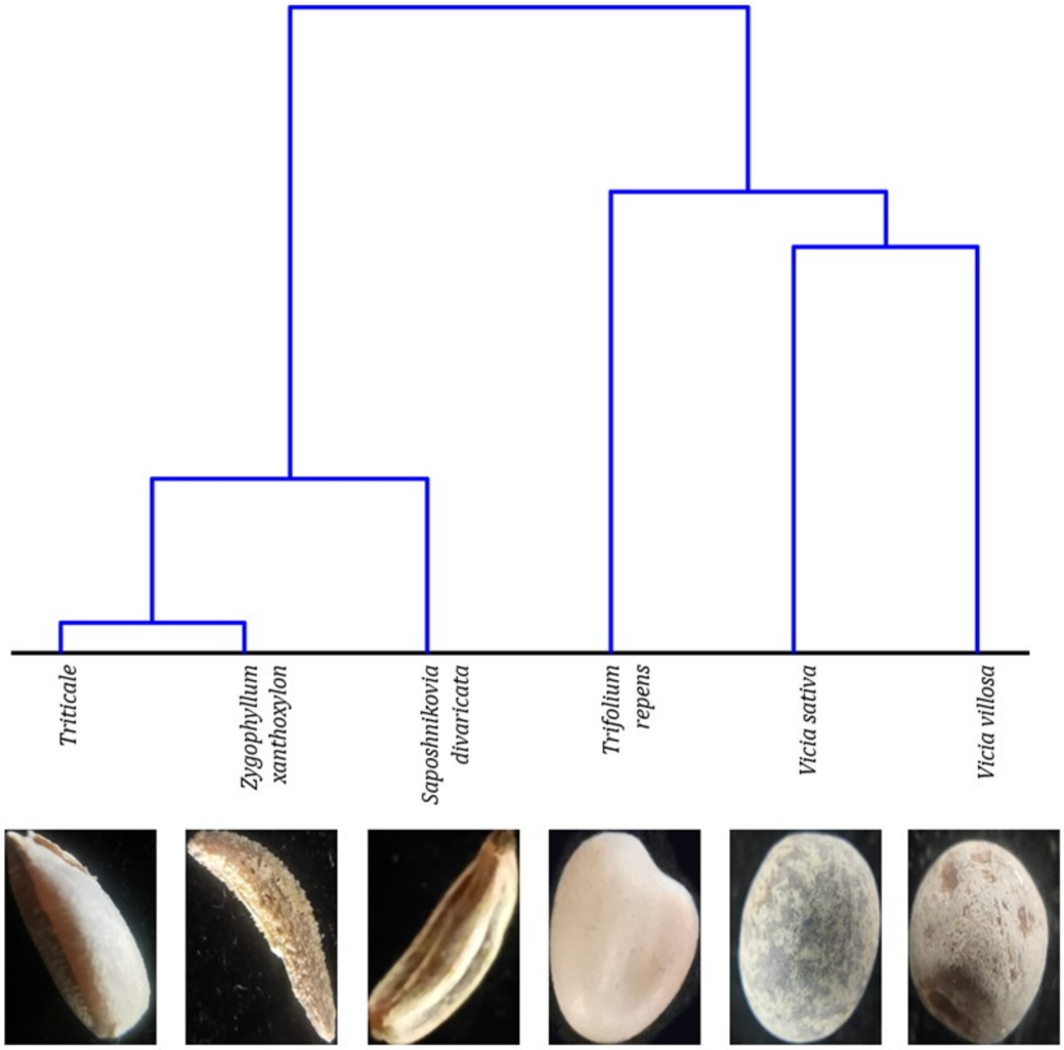

Fig. 6. Dendrogram obtained for another six exemplary seeds species, and representative photos of seeds from these six species.

reflecting strong similarity in their seed characteristics. *Trifolium repens* joins this cluster at a higher dissimilarity threshold, suggesting partial similarity to the Vicia species while maintaining clear species-specific differences. At the highest level of the hierarchy, the left and right clusters merge, indicating that the inter-group dissimilarity is greater than the variation within either group.

MANOVA confirmed highly significant multivariate differences among the six species. Canonical variate analysis showed that interspecific separation was governed primarily by shape-related descriptors. CV1 was dominated by Solidity, Extent and Convex Area, indicating that differences in seed convexity and geometric outline represented the main axis of discrimination. CV2 further captured variation in convexity and proportionality, while CV3 was associated mainly with Major Axis Length and Extent. Texture-based descriptors contributed minimally to all canonical variates.

Pairwise Mahalanobis distances revealed distinct grouping patterns. The closest species pair was *Triticale* and *Zygophyllum xanthoxylon* (d = 10.80), followed by partial similarity between *Vicia sativa* and *Vicia villosa* (d = 14.26). In contrast, *Vicia villosa* was the most divergent taxon, showing the largest separation from *Saposhnikovia divaricata* (d = 35.72).

Fig. 7 presents a heatmap, illustrating the analysis of individual seed features for the second example.

Comparison of this heatmap with the one presented in Fig. 5 leads to several observations. As in the previous case, the features Filled Area and Area appear closely related. However, other pairs of features do not exhibit such clear proximity. Overall, the feature dendrogram in this heatmap differs noticeably from the previous one, indicating that increasing taxonomic diversity alters the similarity structure among descriptors. This reflects changes in how features co-vary across the broader set of species, rather than shifts in their discriminative influence. Nevertheless, when examining the individual columns (representing species), a degree of similarity can still be observed. This indicates that while the discriminative power of features may vary, their behaviour across clusters remains relatively consistent.

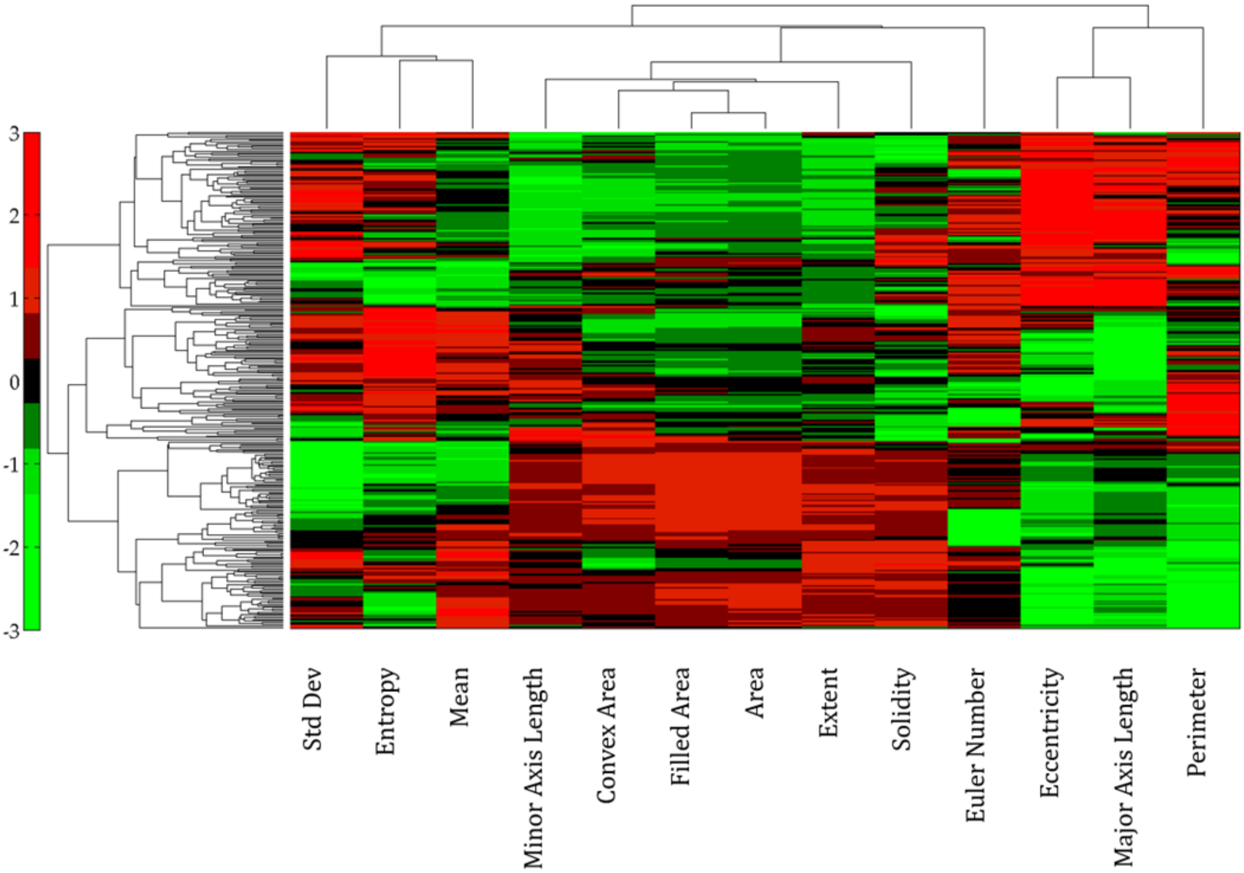

Fig. 7. Hierarchical clustering heatmap of seed features for six exemplary seeds species.

Analysis of differences and similarities at genus level

To present the results of the analysis at the genus level, we included only genera represented by at least two distinct species in our dataset. This criterion ensured sufficient intra-genus variation for meaningful analysis. Applying this rule resulted in the selection of fifteen genera (family indicated in parentheses):

Achnatherum, *Agropyron, Elymus, Puccinellia (*Poaceae*)*, *Apocynum* (Oleanderaceae*)*, *Artemisia (*Compositae*)*, *Astragalus*, *Caragana, Corethrodendron, Lespedeza, Medicago, Vicia (*Leguminosae*)*, *Lycium (*Solanaceae*)*, *Nitraria (*Nitrariaceae*)*, and *Rumex (*Polygonaceae*)*.

The dendrogram generated for these genera is shown in Fig. 8.

This image offers several noteworthy insights. Most strikingly, two clearly separated branches correspond to the genera *Nitraria* and *Rumex*, indicating that these groups possess distinctive sets of traits and are unlikely to be confused with seeds from the other analysed genera. Further examination reveals that *Puccinellia* and *Vicia* also form relatively distinct clusters, suggesting they, too, are set apart by their geometric and textural features.

Two larger clusters are visible: one comprising *Artemisia*, *Medicago*, *Lycium*, and *Lespedeza*, and another containing *Astragalus*, *Corethrodendron*, and *Caragana*. The y-axis values indicate that the differences among genera in the first group are smaller than those in the second. Notably, the first cluster includes genera from three families (Compositae, Leguminosae, and Solanaceae), whereas all genera in the second group belong to the Leguminosae family.

Interestingly, genera like *Elymus* (Poaceae) and *Apocynum* (Oleanderaceae), despite their taxonomic separation, appear closely grouped. Such similarities in nature may reflect convergent adaptations to similar ecological niches or dispersal strategies; however, exploring these aspects lies beyond the scope of this article.

MANOVA revealed highly significant multivariate differences among the 15 genera. Canonical variate analysis demonstrated that interspecific separation was driven primarily by geometric and convexity-related descriptors. CV1 was dominated by Solidity, Extent and Major Axis Length, indicating that seed compactness and elongation represent the main axes of morphological differentiation. CV2 captured additional differences in convexity and outline regularity, whereas CV3 reflected variation in elongation and proportionality. Texture-related features had only marginal contributions across all canonical variates.

Mahalanobis distances showed strong taxonomic structuring. The closest genera were *Agropyron* and *Elymus* (d = 1.74), with *Achnatherum*

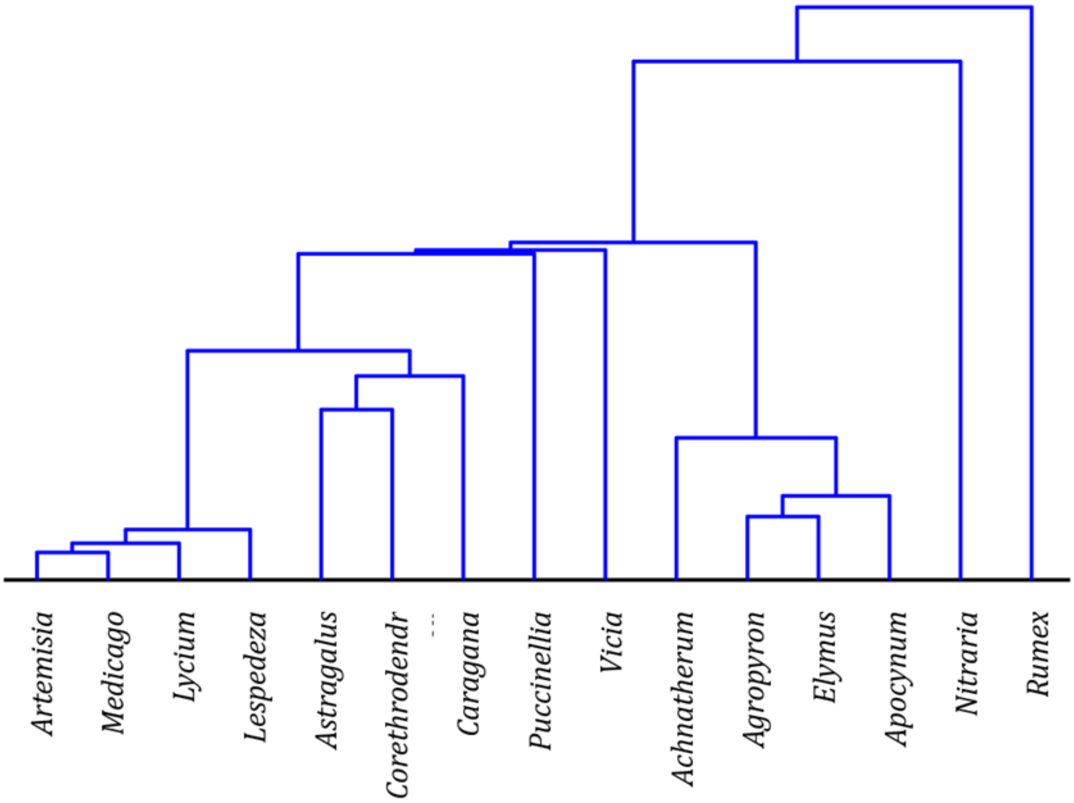

Fig. 8. Hierarchical clustering of plant genera based on seed morphological and textural features.

also clustering closely with them, forming a cohesive Poaceae group. Fabaceae members (*Astragalus, Corethrodendron, Vicia*) also showed moderate similarity. In contrast, *Rumex* was the most distinct genus, exhibiting the largest distances to all others (up to d = 22.31).

The heatmap shown in Fig. 9 provides a visualization of variation in morphological and textural seed traits across all analysed samples.

In this heatmap, clear clustering patterns are observed both among features and observations. Size-related features such as Area, Filled Area, and Convex Area cluster tightly, reflecting their strong similarity and shared measurement characteristics. Textural descriptors (Entropy, Mean, and Std Dev) also form a coherent cluster, indicating that these variables exhibit similar patterns across samples and behave consistently as a group within the dataset.

Analysis of differences and similarities at family level

Fig. 10 shows a dendrogram constructed based on the statistical analysis of feature vectors of seeds grouped by families. 68 species were analyzed, i.e. only those families that were represented by at least two species were selected.

The clustering reveals several notable patterns. Leguminosae, Lamiaceae, and Amaranthaceae form a closely associated group, indicating a high degree of similarity in seed features such as shape, surface texture, and compactness. Similarly, Poaceae, Brassicaceae, and Solanaceae are grouped together, suggesting shared morphological traits, possibly related to their more regular or elongated seed forms.

At the other end of the dendrogram, Polygonaceae and Nitrariaceae are positioned as the most distinct families. Their separation from all other groups at a higher linkage distance suggests a unique combination of features, which may include greater variance in seed eccentricity, perimeter, or textural complexity.

Interestingly, Elaeagnaceae and Oleanderaceae also form a discrete cluster that joins the rest of the dendrogram only at a higher level, implying partial but limited similarity with other families, particularly those in the intermediate group such as Ranunculaceae and Compositae.

MANOVA confirmed highly significant multivariate differences among the 12 plant families. Canonical variate analysis demonstrated that the primary axes of discrimination were governed by shape-related descriptors. CV1 was dominated by Solidity, ConvexArea and MajorAxisLength, indicating that differences in seed compactness and elongation represent the principal source of family-level separation. CV2 and CV3 reflected additional variation in convexity and proportionality of the seeds. Texture-based descriptors contributed minimally to the canonical structure.

Mahalanobis distances revealed several coherent morphological clusters. The closest families were Leguminosae–Lamiaceae (d = 1.19), Leguminosae–Solanaceae (d = 1.65) and Poaceae–

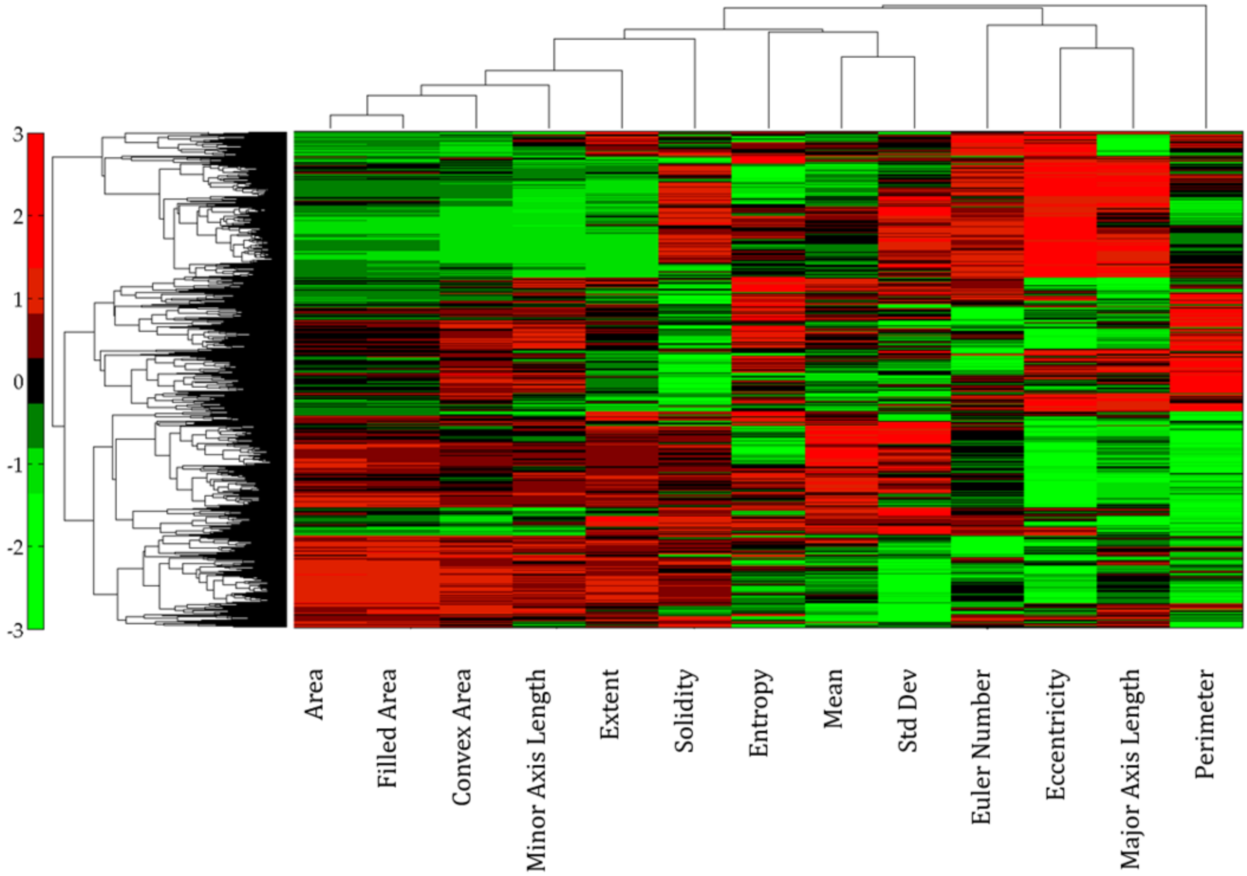

Fig. 9. Hierarchical clustering heatmap of seed features for fifteen seeds genera.

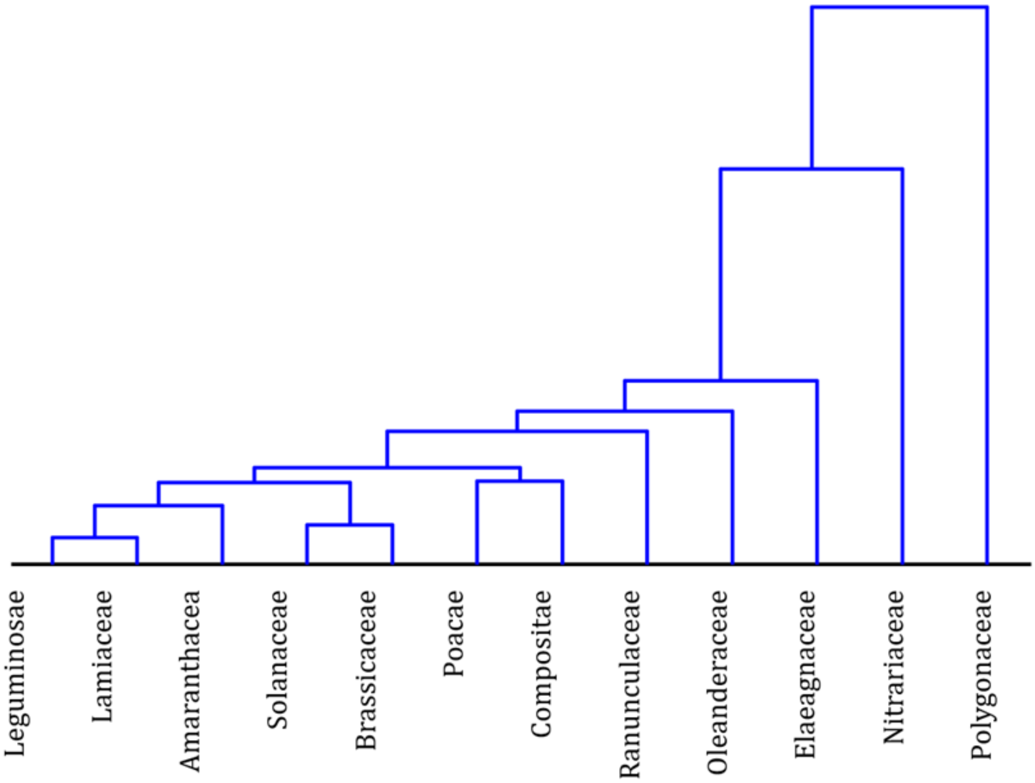

Fig. 10. Hierarchical clustering of seeds features in plant families.

Compositae (d = 1.67). In contrast, Polygonaceae exhibited the greatest divergence, with distances exceeding 16 to several other families, followed by Oleanderaceae and Nitrariaceae.

The heatmap shown in Fig. 11 provides a visualization of variation in morphological and textural seed traits across all analyzed samples for families clustering.

Features dendrogram for heatmap in Fig. 11 exhibits similarities to the one in Fig. 9, suggesting that analogous features are fundamental to the analysis at both the family and genus levels.

To compare how feature similarity patterns vary across taxonomic levels, Fig. 12 presents dendrograms of seed descriptors generated at the species, genus, and family levels, based on the heatmaps shown in Figs. 5, 7, 9, and 11.

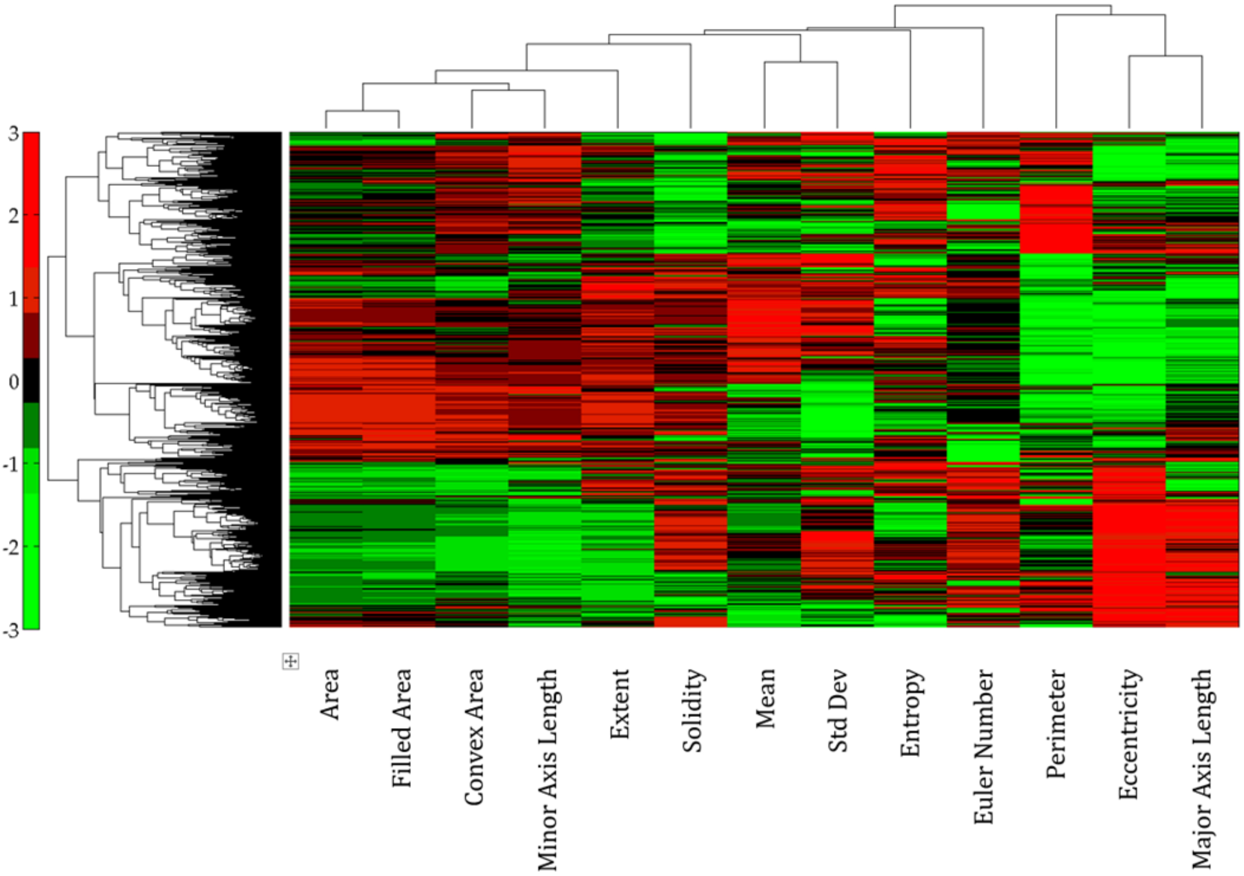

Fig. 11. Heatmap obtained for features of seeds grouped by families.

Across all taxonomic levels, some consistent groupings of features can be observed - for instance, Area and Filled Area remain closely clustered, which reflects their intrinsic relationship. Likewise, Major Axis Length, Perimeter, and Eccentricity frequently form a group, suggesting that these features jointly characterize shape elongation and contour complexity.

At the species level, the dendrograms show a higher degree of variability among feature groupings, likely due to the distinct morphological characteristics specific to individual species. This observation is confirmed by the fact that the dendrograms for two different species sets show noticeable differences.

At the genus level, clustering becomes more consistent, with a more balanced distribution of groups. The proximity of Convex Area and Minor Axis Length suggests they capture related aspects of seed morphology within genera, though this relationship appears more variable at the species level.

This may indicate that certain features become less variable or display more generalized patterns as the taxonomic resolution decreases, with broad morphological trends dominating at the family level.

Generally, the comparison suggests that while certain morphological features remain robust indicators across taxonomic levels, others show varying degrees of influence depending on the biological diversity being considered. This emphasizes the need for careful feature selection tailored to the intended comparison level.

Across all taxonomic levels analysed (species, genera and families), multivariate separation was consistently governed by the same subset of shape -related descriptors, primarily Solidity, Extent,

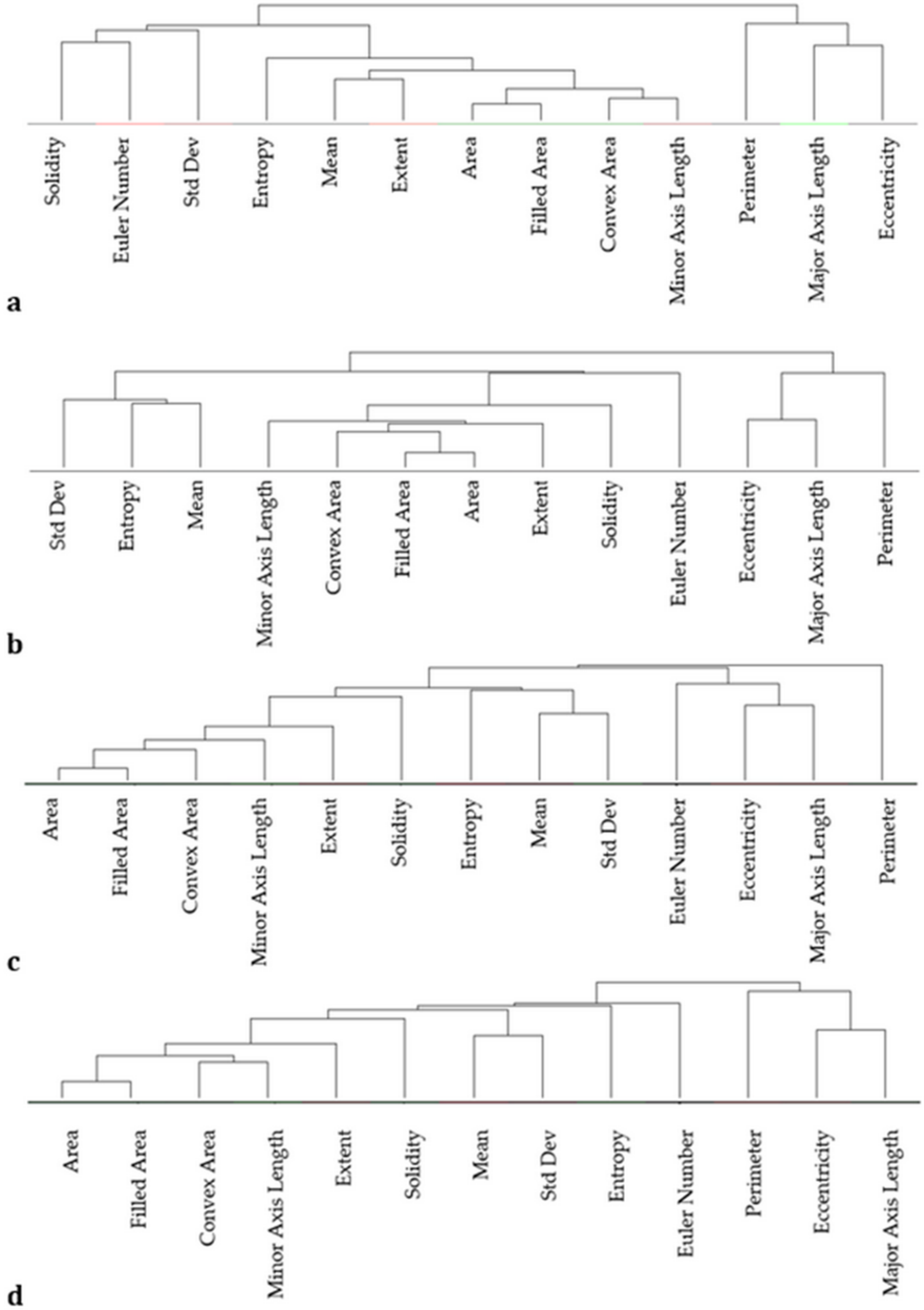

Fig. 12. Comparison of dendrograms of seeds features obtained at species (a and b, for the first and second example, respectively), genus (c) and family (d) levels.

ConvexArea and MajorAxisLength. Texture-based features played only a minor role irrespective of the taxonomic rank. This suggests that seed outline geometry provides a robust and scale-independent source of taxonomic signal, whereas texture contributes only marginally to interspecific and higher-level differentiation.

In comparison with previous studies that analysed seed morphology and texture for species classification (e.g., Djoulde, 2024; Espinosa-Roldán et al., 2024; Ermiş et al., 2025), the present study adopts a unified set of fundamental geometric and textural features across multiple taxonomic levels (species, genus, and family). While prior work often focused on individual species or task-specific descriptors, our approach enables both intra- and inter-taxon comparisons, revealing which features consistently discriminate among taxa. Notably, the analyses show that some descriptors, such as Area, Filled Area, Major Axis Length, and Eccentricity, maintain high discriminative power across taxonomic scales, whereas others vary in influence depending on the level of taxonomic resolution. This demonstrates the potential of simple, explainable features to provide biologically meaningful insights, support automated classification, and complement more complex, black-box approaches. Overall, these findings extend previous knowledge by identifying robust descriptors applicable across multiple levels and highlighting patterns of seed morphology that were not previously characterized in a comparative framework.

Conclusions

This study demonstrates that fundamental geometric and textural descriptors extracted from seed images can effectively capture biologically meaningful patterns and reflect taxonomic relationships at the species, genus and family levels. The combined use of MANOVA, canonical variate analysis and Mahalanobis distances confirmed highly significant multivariate differences across all taxonomic ranks, with most separation governed by a consistent subset of shape-related features. Texture -based descriptors played only a minor role, indicating that seed outline geometry provides a robust and scale-independent source of taxonomic signal.

Hierarchical clustering and heatmaps further revealed coherent similarity structures among taxa and among descriptors, supporting intuitive groupings and highlighting redundancies between some shape measures. Although exploratory in nature, these visualizations complemented the inferential analyses by aiding interpretation of multivariate patterns and illustrating consistent feature behaviour across taxonomic levels.

Overall, the results indicate that simple, interpretable descriptors offer a reliable basis for distinguishing plant taxa and have strong potential for integration into automated identification systems. By relying on explainable image-derived metrics rather than black-box models, the proposed approach supports transparent taxonomic inference and provides a practical framework for biodiversity research and applied seed recognition.

Limitations

One important limitation of this study concerns the quality of some seed images used for analysis. Acquisition artifacts such as blurring, overexposure, or uneven lighting can influence the calculated shape and texture descriptors, potentially introducing noise into the clustering results. Consequently, interpretations based on individual images should be made with caution, and firm conclusions should not rely on examples affected by acquisition artifacts.

However, the analysis was based on a set of 13 descriptors, which reduces the impact of noise in individual features. In cases where certain descriptors are affected by image quality, their influence is mitigated by the remaining unaffected features, allowing the clustering algorithm to rely on the overall descriptor set rather than isolated values. This design ensures that the method remains robust even when minor imperfections occur in the input data.

References

Barbierato, E., & Gatti, A. (2024). The challenges of machine learning: A critical review. Electronics, 13(2), 416.

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. Machine Learning and Knowledge Extraction, 3(4), 966–989.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. Remote Sensing, 13 (22), 4712.

Chen, Z., Fan, W., Luo, Z., & Guo, B. (2022). Soybean seed counting and broken seed recognition based on image sequence of falling seeds. Computers and Electronics in Agriculture, 196, 106870.

Cho, M., & Martinez, W. L. (2014). Statistics in MATLAB: A Primer. CRC Press: Boca Raton, USA.

Djoulde, K., Ousman, B., Hamadjam, A., Bitjoka, L., & Tchiegang, C. (2024). Classification of pepper seeds by machine learning using color filter array images. Journal of Imaging, 10(2), 41.

Engle, S., Whalen, S., Joshi, A., & Pollard, K. S. (2017). Unboxing cluster heatmaps. BMC Bioinformatics, 18, 1–15.

Ermiş, S., Ercan, U., Kabaş, A., Kabaş, Ö., & Moiceanu, G. (2025). Machine learning-based morphological classification and diversity analysis of ornamental pumpkin seeds. Foods, 14(9), 1498.

Eryigit, R., & Tugrul, B. (2021). Performance of various deep -learning networks in the seed classification problem. Symmetry, 13(10), 1892.

Espinosa-Roldán, F. E., Rodríguez-Lorenzo, J. L., Martín-Gómez, J. J., Tocino, Á., Ruiz Martínez, V., Remón Elola, A., Cabello Sáenz de Santamaría, F., Martínez de Toda, F., Cervantes, E., & Muñoz-Organero, G. (2024). Morphometric analysis of grape seeds: Looking for the origin of Spanish cultivars. Seeds, 3(3), 286–310.

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2003). Digital Image Processing Using MATLAB. Prentice Hall: New Jersey, USA.

Islam, T., Sarker, T. T., Ahmed, K. R., & Lakhssassi, N. (2024). Detection and classification of cannabis seeds using RetinaNet and Faster R-CNN. Seeds, 3(3), 456–478.

Kalicka, R., & Lipiński, S. (2010). A fast method of separation of the noisy background from the head-cross section in the sequence of MRI scans. Biocybernetics and Biomedical Engineering, 30(2), 15–27.

Krohn, J., Beyleveld, G., & Bassens, A. (2019). Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence. Addison-Wesley Professional: Boston, USA.

Kumar, V., Aydav, P. S. S., & Minz, S. (2024). Crop Seeds Classification Using Traditional Machine Learning and Deep Learning Techniques: A Comprehensive Survey. SN Computer Science, 5(8), 1031.

Lipiński, A. J., & Lipiński, S. (2020). Binarizing water sensitive papers–how to assess the coverage area properly?. Crop Protection, 127, 104949.

Loddo, A., Loddo, M., & Di Ruberto, C. (2021). A novel deep learning based approach for seed image classification and retrieval. Computers and Electronics in Agriculture, 187, 106269.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62–66.

Pavlopoulos, G. A., Soldatos, T. G., Barbosa-Silva, A., & Schneider, R. (2010). A reference guide for tree analysis and visualization. BioData Mining, 3, 1–24.

Rajalakshmi, R., Faizal, S., Sivasankaran, S., & Geetha, R. (2024). RiceSeedNet: Rice seed variety identification using deep neural network. Journal of Agriculture and Food Research, 16, 101062.

Reyes-Aldasoro, C. C. (2015). Biomedical Image Analysis Recipes in MATLAB: For Life Scientists and Engineers. John Wiley & Sons: Oxford, UK.

Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging, 13(1), 146–165.

Szandała, T. (2023). Unlocking the black box of CNNs: Visualising the decision-making process with PRISM. Information Sciences, 642, 119162.

Taheri-Garavand, A., Nasiri, A., Fanourakis, D., Fatahi, S., Omid, M., & Nikoloudakis, N. (2021). Automated in situ seed variety identification via deep learning: A case study in chickpea. Plants, 10(7), 1406.

Taye, M. M. (2021). Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. Computation, 11(3), 52.

Wang, L., & Wang, L. (2021). Variety identification model for maize seeds using hyperspectral pixel-level information combined with convolutional neural network. National Remote Sensing Bulletin, 25(11), 2234–2244.

Wiesnerova, D & Wiesner, Ivo (2008). Computer image analysis of seed shape and seed color for flax cultivar description. Computers and Electronics in Agriculture, 61 (2), 126–135.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. The American Statistician, 63(2), 179–184.

Yasar, A. (2024). Analysis of selected deep features with CNN-SVM-based for bread wheat seed classification. European Food Research and Technology, 250(6), 1551–1561.

Yuan, M., Lv, N., Dong, Y., Hu, X., Lu, F., Zhan, K., ... Xie, Y. (2024). A dataset for fine-grained seed recognition. Scientific Data, 11(1), 344.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99.