Erciyes University Journal of Institute Of Science and Technology

Research Article
10.65520/erciyesfen.1980963

Imprint:
Volume: 42(3)
Year: 2026
Page: 495-522

Sevim Akçura[a*]

[a] Dr., Çanakkale Onsekiz Mart University, sevimakcura@yahoo.com

[*] Corresponding Author

Received: 6/28/2026
Accepted: 10/21/2026

Citation:
Sevim Akçura (2026). Determining Phenotypic Relationships among Flax (*Linum usitatissimum* L.) Genotypes Using Machine Learning and Deep Learning Guided by Phenotypic Core-Based Binary Seed Image Classification. *Erciyes University Journal of Institute Of Science and Technology*, 42(3), 495-522. https://doi.org/10.65520/erciyesfen.*1980963*

Screened by
iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

Determining Phenotypic Relationships among Flax (*Linum usitatissimum* L.) Genotypes Using Machine Learning and Deep Learning Guided by Phenotypic Core-Based Binary Seed Image Classification

Abstract

Background/Aim: Accurate identification of flax (*Linum usitatissimum* L.) genotypes is essential for germplasm management, breeding programmes and seed certification. Although image-based phenotyping and deep learning have shown considerable potential for genotype classification, the relationships between phenotypic similarity, within-genotype variation and classification performance remain poorly understood. This study aimed to investigate binary genotype classification using phenotypic core seed sets and to examine the relationships among deep feature space, phenotypic distance and classification performance.

Methods: Single-seed images representing five flax genotypes were used to construct phenotypic core datasets by selecting representative individuals from each genotype. Binary classification was performed using four convolutional neural network architectures. Deep feature space extracted by DenseNet121 was further used to quantify phenotypic distances, evaluate within-genotype variation, construct phenotypic similarity networks and visualise genotype relationships using hierarchical clustering and UMAP.

Results: Significant differences were detected among genotypes in morphological, colour and texture characteristics ($P < 0.001$). Binary classification performance varied among genotype pairs, with F1 scores ranging from 0.706 to 0.951. DenseNet121 achieved the highest performance in six of the ten binary comparisons, while MobileNetV2 produced the highest overall F1 score (0.951). Phenotypic distances calculated in the deep feature space showed a significant positive correlation with CNN classification performance (Pearson $r = 0.710$, $P = 0.021$), whereas no significant relationship was observed for conventional machine learning models. Phenotypic similarity networks, hierarchical clustering and UMAP projections consistently visualised phenotypic relationships among genotypes and highlighted differences in within-genotype homogeneity.

Conclusions: The findings demonstrate that image-based deep phenotyping can provide biologically interpretable information beyond classification accuracy. Integrating phenotypic core selection with deep feature space analysis offers a complementary framework for evaluating genotype relationships, within-genotype variation and representative germplasm, supporting future applications in flax breeding and digital phenotyping.

Keywords: Digital phenotyping, phenotypic core, deep feature space, phenotypic distance, germplasm characterization, seed phenotyping

Fenotipik Çekirdek Rehberli İkili Tohum Görüntüsü Sınıflandırması ile Makine Öğrenmesi ve Derin Öğrenme Kullanılarak Keten (*Linum usitatissimum* L.) Genotipleri Arasındaki Fenotipik İlişkilerin Belirlenmesi

Öz

Amaç: Keten (*Linum usitatissimum* L.) genotiplerinin doğru şekilde tanımlanması; genetik kaynakların etkin yönetimi, ıslah programlarının yürütülmesi ve tohum sertifikasyonu açısından büyük önem taşımaktadır. Görüntü tabanlı fenotipleme ve derin öğrenme yöntemleri genotip sınıflandırmasında başarılı sonuçlar vermesine rağmen, fenotipik benzerlik, genotip içi varyasyon ve sınıflandırma başarısı arasındaki ilişkiler henüz yeterince açıklığa kavuşturulamamıştır. Bu çalışmada, fenotipik çekirdek tohum veri setleri kullanılarak ikili genotip sınıflandırması gerçekleştirilmiş ve derin özellik uzayı, fenotipik uzaklık ile sınıflandırma performansı arasındaki ilişkilerin ortaya konulması amaçlanmıştır.

Yöntemler: Beş farklı keten genotipine ait tek tohum görüntüleri kullanılarak, her genotipi en iyi temsil eden bireylerin seçilmesiyle fenotipik çekirdek veri setleri oluşturulmuştur. İkili genotip sınıflandırmaları dört farklı evrişimsel sinir ağı mimarisi kullanılarak gerçekleştirilmiştir. Ayrıca, DenseNet121 modelinden elde edilen derin özellik uzayı kullanılarak genotipler arasındaki fenotipik uzaklıklar hesaplanmış, genotip içi varyasyon değerlendirilmiş, fenotipik benzerlik ağları oluşturulmuş ve genotipler arasındaki ilişkiler hiyerarşik kümeleme ile UMAP yöntemleri kullanılarak görselleştirilmiştir.

Bulgular: Genotipler arasında morfolojik, renk ve doku özellikleri bakımından istatistiksel olarak anlamlı farklılıklar belirlenmiştir ($P < 0,001$). İkili sınıflandırma performansı genotip çiftlerine göre değişmiş ve F1 skorları 0,706 ile 0,951 arasında bulunmuştur. DenseNet121 modeli on ikili karşılaştırmanın altısında en yüksek performansı gösterirken, en yüksek genel F1 skoru (0,951) MobileNetV2 modeli ile elde edilmiştir. Derin özellik uzayında hesaplanan fenotipik uzaklıklar ile evrişimsel sinir ağı modellerinin sınıflandırma performansı arasında anlamlı ve pozitif bir ilişki saptanmıştır (Pearson $r = 0,710$; $P = 0,021$). Buna karşılık, geleneksel makine öğrenmesi modellerinde benzer bir ilişki belirlenmemiştir. Fenotipik benzerlik ağları, hiyerarşik kümeleme ve UMAP analizleri, genotipler arasındaki fenotipik ilişkileri tutarlı biçimde görselleştirmiş ve genotip içi homojenlik düzeylerindeki farklılıkları açıkça göstermiştir.

Sonuç: Araştırma bulguları, görüntü tabanlı derin fenotipleme yaklaşımının yalnızca yüksek sınıflandırma doğruluğu sağlamakla kalmayıp, aynı zamanda genotipler arasındaki biyolojik ilişkilerin yorumlanmasına da önemli katkılar sunduğunu göstermektedir. Fenotipik çekirdek seçiminin derin özellik uzayı analizi ile birlikte kullanılması; genotipler arasındaki fenotipik ilişkilerin, genotip içi varyasyonun ve temsil edici genetik materyalin değerlendirilmesine yönelik tamamlayıcı ve güçlü bir yaklaşım sunmaktadır. Bu yaklaşımın keten ıslahı, genetik kaynakların karakterizasyonu ve dijital fenotipleme çalışmalarında gelecekte önemli uygulama alanları bulabileceği değerlendirilmektedir.

Anahtar kelimeler: Dijital fenotipleme, fenotipik çekirdek, derin özellik uzayı, fenotipik uzaklık, genetik kaynakların karakterizasyonu, tohum fenotiplemesi

❋ ❋ ❋

1. Introduction

Flax (*Linum usitatissimum* L.) is an economically important oil- and fiber-producing crop whose genetic resources are relevant to breeding, seed quality and varietal-purity management [1-6]. Reliable genotype characterization is therefore important, but conventional morphological, biochemical and molecular approaches can be subjective, destructive, time-consuming or infrastructure-intensive. Image-based seed phenotyping provides a rapid, non-destructive alternative by quantifying seed morphology, color and surface texture under standardized acquisition conditions [7-18].

Recent advances in image-based phenotyping, machine learning, and deep learning have expanded the possibilities for seed characterization and cultivar discrimination across various crop species, including flax [15, 16, 19–27]. Akbarpour et al. [22] investigated deep-learning-based maize seed variety classification, while Wei et al. [23] combined multispectral seed imaging with machine-learning algorithms for varietal discrimination. Lim et al. [25] integrated seed morphology with

CC: BY-NC-ND

geographic origin and genetic structure, whereas Lipiński [26] examined morphological and textural characteristics for assessing seed similarity. Wang et al. [27] further investigated maize seed-coat phenotypes using detailed morphological, colorimetric, and textural descriptors. Collectively, these studies demonstrate the increasing use of quantitative image-derived traits and computational approaches for seed characterization and classification. Nevertheless, the relationships among representative-sample selection, within-genotype phenotypic dispersion, deep feature-space distances, and pairwise classification performance remain insufficiently characterized, particularly in flax germplasm.

Another important but only partially investigated topic is within-genotype phenotypic variation. Natural differences can exist among seeds belonging to the same genotype, and some individuals may constitute outliers that do not represent the general phenotypic structure of the population [1, 2]. This situation can introduce noise into image-based classification models, affecting the decision boundaries and making it more difficult to evaluate the true phenotypic relationships among genotypes. Consequently, selecting the individuals that best represent a genotype can contribute to developing more stable and biologically more meaningful models. The core-collection approach, used for many years in genetic resource management, similarly shows that selecting highly representative subsets from large collections can increase the efficiency and interpretability of analyses [3-5]. Nevertheless, studies in which phenotypic core samples are constructed in image-based seed phenotyping and in which the effect of this approach on both classification performance and phenotypic relationships is evaluated jointly are rather limited.

Accordingly, this study investigated whether representative phenotypic core seed sets can support interpretable pairwise discrimination among five flax genotypes. Three questions were addressed: (i) do the genotypes differ in image-derived morphological, colour and texture traits; (ii) does pairwise classification performance vary systematically with phenotypic separation in a common deep feature space; and (iii) does within-genotype dispersion provide complementary information about phenotypic homogeneity? We hypothesized that genotype pairs separated by larger DenseNet121 feature-space distances would, on average, be more readily discriminated by CNN models, while acknowledging that distance is not expected to determine performance perfectly for every pair. The phenotypic-core approach was used as a representative-subset strategy rather than as a claim of genetic core-collection equivalence.

2. Materials and Methods

2.1. Plant material

Five flax (*Linum usitatissimum* L.) genotypes (BeyazGelin, Culbert, EBYC01, K0902, and K1104) were used as plant material in this study. The seed material was obtained from the Trakya Agricultural Research Institute. Three genotypes (EBYC01, K0902, and K1104) were advanced breeding lines developed within the Turkish National Oilseed Flax Breeding Program, whereas BeyazGelin and Culbert were included as control cultivars. The advanced breeding lines were evaluated alongside the control cultivars under the agroecological conditions of the Southern Marmara Region to assess their regional adaptation and performance prior to broader multi-environment evaluation.

The research material used for seed phenotyping was obtained from field experiments conducted during the 2019–2020 and 2020–2021 growing seasons at the experimental fields of the Farmer Training Branch Directorate (BAÇEM), operating under the Department of Rural Services of the Balıkesir Metropolitan Municipality, Türkiye. The experiments were established in a randomized complete block design with three replications. In both years, sowing was carried out in the first week of March in 0.8 × 5.0 m (4.0 m^2) plots at a density of 1,000 seeds m^{-2}. The experiments were conducted under rain-fed conditions, with no supplementary irrigation. Meteorological data for the experimental site were obtained from the Burhaniye Meteorological Station (Station No. 17722) of the Turkish State Meteorological Service. During the March–June growing period, monthly mean temperatures were 11.8, 14.0, 19.9, and 23.6 °C in the 2019–2020 season and 9.1, 14.2, 20.8, and 24.1 °C in the 2020–2021 season, respectively. The corresponding monthly precipitation totals were 14.6, 32.6, 57.8, and 24.4 mm in 2019–2020 and 93.0, 50.4, 14.6, and 26.0 mm in 2020–2021. Accordingly, cumulative

CC: BY-NC-ND

precipitation from March to June was 129.4 mm in 2019–2020 and 184.0 mm in 2020–2021. Although the seasonal temperature pattern was broadly similar between the two growing seasons, both the amount and within-season distribution of precipitation differed considerably. Precipitation was substantially higher in March during the 2020–2021 season, whereas May received considerably more precipitation during the 2019–2020 season.

The experimental soil was classified as slightly alkaline, with a pH of 7.4. It had low salinity (<4 dS m^{-1}), very low organic matter content (0.7%), and a sandy loam texture. The soil was deficient in available phosphorus, potassium, and total nitrogen, whereas most of the evaluated micronutrients were present at levels considered adequate for crop production. At sowing, diammonium phosphate (DAP) was applied as basal fertilizer, supplying 80 kg P_2O_5 ha^{-1} and approximately 30 kg N ha^{-1}. Before flowering, an additional 50 kg N ha^{-1} was applied as a top-dressing using ammonium nitrate fertilizer containing 33% N. Thus, the total nitrogen application was approximately 80 kg N ha^{-1}.

To minimize edge effects at harvest, the middle two rows of each plot were harvested; after the resulting seeds had been threshed by hand, they were stored in a cold room at approximately 4°C until analysis. Only fully mature, healthy, and physically undamaged seeds were used in the image analyses; broken, malformed, mechanically damaged seeds or those showing pronounced color defects were excluded from the evaluation.

For the image-based analyses, 84 independent seeds were randomly selected from each of the three field replications in each growing season. Thus, 252 new physical seeds were sampled per genotype in each year (84 seeds × 3 replications), giving 504 independent physical seeds per genotype across the two years (252 seeds $year^{-1}$ × 2 years) and 2,520 single-seed images in total for the five genotypes. The seeds sampled in one growing season were not reused in the other growing season. Each physical seed was imaged once only and constituted one independent image-level observation. Accordingly, the two-year design does not represent repeated imaging of the same seeds.

2.2. Image acquisition and dataset preparation

The overall experimental workflow of the study is summarized in Figure 1. The analysis process consisted, respectively, of image acquisition, image processing and feature extraction, construction of phenotypic core datasets, development of classification models, and feature-space analyses.

Seed images were acquired at 1200 dpi optical resolution using an Epson Perfection V39 flatbed scanner (Seiko Epson Corp., Japan). A single standardized acquisition protocol was used throughout: the scanner lid remained closed, automatic colour correction, sharpening and contrast enhancement were disabled, and the same resolution and scanner settings were retained for all samples. Images were saved as PNG files in the RGB colour space. Representative original multi-seed scans, quality-control examples, segmentation masks and single-seed regions of interest are provided in Figure 1 to document the acquisition and preprocessing workflow. These details are reported to improve reproducibility; transferability to other scanners or field/mobile imaging systems was not assumed.

In each scan, the seeds were positioned so that they did not touch one another, and the images were saved in PNG format in the RGB colour space. Each seed was separated from the background using OpenCV-based image-processing algorithms, and independent single-seed images were created. For each CNN architecture, the images were rescaled to match the input size of the relevant model. Thus, all images were subjected to the same pre-processing steps, minimizing the technical variation that might arise from imaging conditions.

At the end of the imaging process, the single-seed images obtained for each genotype were merged into a common data pool and were used in the subsequent image processing, phenotypic core construction, machine-learning, deep-learning and feature-space analyses. Each image was cropped to contain only a single seed, and the file names were standardized on a genotype basis.

CC: BY-NC-ND

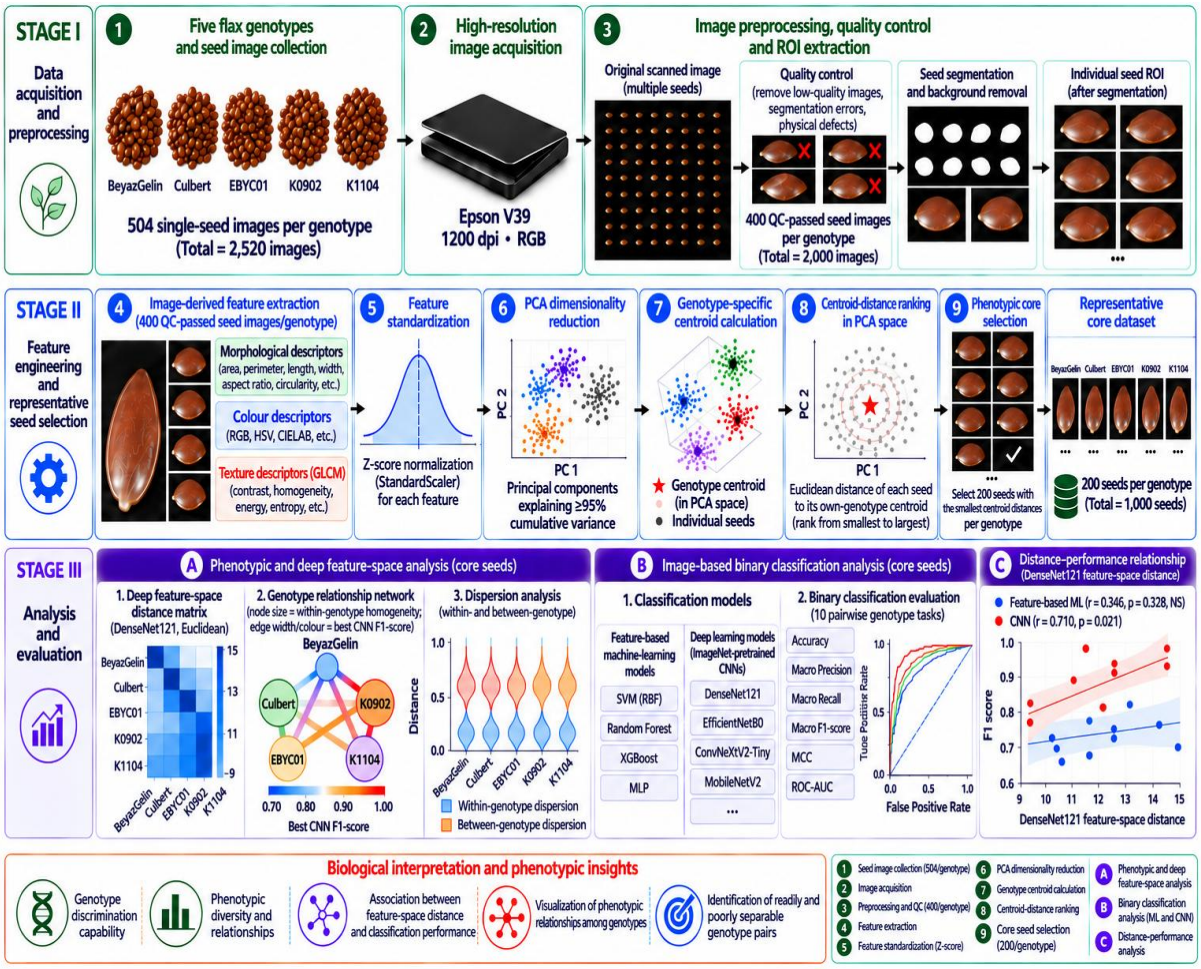

Figure 1. Workflow of the proposed representative seed selection, feature-based machine learning, and binary deep learning framework.

2.3. Image processing and feature extraction

Image processing and feature extraction were carried out by following a predefined, standard image-processing workflow to obtain comparable and reproducible measurements across images (Figure 1). In the first stage, the seeds in each scanned image were separated from the background, and each seed was defined as an independent region of interest (ROI) containing a single seed. After ROI determination, each seed was cropped individually to create the independent images to be used in the subsequent analyses.

Segmentation was performed by applying thresholding to the images converted to greyscale using OpenCV and by generating a binary mask. As a result of this process, only the pixels representing the seed were included in the analyses, while the pixels belonging to the background were excluded from all measurements. In this way, the morphological, color and texture features were ensured to represent only the seed surface, and errors that might arise from imaging conditions were minimized.

Three different feature groups were obtained from the segmented images. The morphological features comprised seed area, perimeter, width, height, major axis length, minor axis length, equivalent diameter, aspect ratio, circularity, extent and solidity. Taking the optical resolution of the scanner into account, the pixel-based measurements were converted into physical units. In this context, one pixel length was taken as 0.02117 mm and one pixel area as 0.000448 mm^2, and all morphological measurements were reported in real length and area units.

CC: BY-NC-ND

Color features were obtained from the RGB, HSV and CIELAB color spaces. In the RGB color space, the mean values of the red (R), green (G) and blue (B) components were calculated, while in the HSV color space the hue (H), saturation (S) and value (V) variables were determined. In addition, the lightness (L*), a* and b* coordinates were calculated from the CIELAB color space, enabling the color differences among genotypes to be evaluated quantitatively. Texture features were calculated using the Gray Level Co-occurrence Matrix (GLCM) approach. GLCM features were calculated for four different directions (0°, 45°, 90° and 135°), and the mean values were used in the analyses [6]. The GLCM analyses yielded the contrast, dissimilarity, homogeneity, energy, correlation, angular second moment (ASM) and entropy variables. By quantitatively describing the tone distribution and the spatial pixel relationships on the seed surface, these parameters were used to evaluate the surface differences among genotypes. During image processing, samples with incomplete segmentation, missing ROIs or non-computable features were removed from the dataset. Variables that do not represent the phenotypic character, such as image file information, were not included in the analyses. The morphological, color and texture features obtained were used in the subsequent phenotypic core construction process, machine-learning analyses and deep-learning-based feature-space evaluations.

2.4. Phenotypic core dataset construction

The main methodological approach of this study is the construction, prior to the model-development stage, of phenotypic core datasets composed of the individuals that best represent each genotype (Figure 1). In conventional image-based classification studies, all images are used directly in model training. In contrast, in this study, rather than eliminating the natural within-genotype variation, a sample-selection approach based on identifying the individuals that represent the typical phenotypic structure of the population was adopted. This approach was adapted from the principles of constructing a representative core subset used in genetic resource studies [3, 4]. The aim was thus to reduce the effect of outliers and atypical phenotypes on the classification models, to develop more stable models and to evaluate the phenotypic relationships among genotypes more reliably.

In the first stage, a total of 504 single-seed images per genotype were evaluated, and samples with low image quality, segmentation errors or physical defects were removed during the quality-control stage. As a result of this process, 400 images per genotype were included in the analyses. The morphological, colour and GLCM-based texture features obtained from each seed were then combined into a common feature matrix and standardized using the StandardScaler method. The dimensionality of the feature space was reduced by Principal Component Analysis (PCA) using the principal components that explained at least 95% of the cumulative variance. The same number of principal components, explaining at least 95% of the cumulative variance, was used for all genotypes. In the resulting PCA feature space, the geometric centroid of each genotype was calculated as the arithmetic mean of the PCA coordinates of the samples belonging to that genotype. With this method, the influence of samples that deviated markedly from the center of the phenotypic distribution was reduced, while the aim was to retain the individuals representing the dominant phenotypic structure of each genotype. Next, the Euclidean distance between the standardized PCA coordinates of each seed and the coordinates of its own genotype center was calculated. This distance was treated as a quantitative indicator of the degree to which an individual represented the typical phenotypic structure of the relevant genotype, and the 200 seeds with the smallest distance to the center were selected to form the phenotypic core dataset for each genotype. Samples with greater distances from the centre were not removed from the dataset; they were simply not included in the phenotypic core dataset. The phenotypic core datasets thus constructed for each genotype were used as the common database in all subsequent analyses of the study. The 200-seed core (50% of the 400 QC-passed images per genotype) was specified as a pragmatic balance between retaining a substantial number of observations and restricting the analysis to individuals nearest the genotype-specific phenotypic centroid. This 50% threshold was not selected through post hoc optimization against classification performance; therefore, it should be regarded as an operational design choice rather than an empirically optimal core size. Sensitivity analyses across alternative core proportions are warranted in future studies.

CC: BY-NC-ND

The feature-based machine-learning classifications, deep-learning classifications, permutation-importance analyses, deep feature-space evaluations, phenotypic distance calculations, within-genotype dispersion analyses, phenotypic similarity networks, hierarchical clustering and UMAP visualisations were all carried out using the same phenotypic core dataset [7]. This common reference dataset increased comparability among the complementary analyses.

2.5. Feature-based machine learning models

Using the phenotypic core datasets, the distinguishability among the five flax genotypes was evaluated with feature-based machine-learning approaches. Because the main aim of the study was to examine in detail the level of phenotypic similarity between genotype pairs, all possible pairwise combinations that could be formed among the five genotypes were evaluated, and a total of ten independent binary classification problems were defined. Each genotype pair was treated as an independent classification problem, and the same model-development and evaluation strategy was applied in all comparisons. The effect of different levels of phenotypic similarity on classification performance could thus be compared directly.

Each seed was represented by a numerical feature vector composed of morphological, colour and GLCM-based texture features, and all machine-learning models were trained on this feature matrix. Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost; XGB) and Multilayer Perceptron (MLP) algorithms were used in the analyses [8-10]. These algorithms were preferred because of their ability to model non-linear decision boundaries, their capacity to adapt to different data structures and their widespread use in image-based phenotyping studies.

Before model training, all variables were standardized using StandardScaler. No separate feature-selection algorithm was applied before classification: the complete predefined set of morphological, colour and GLCM texture descriptors was retained for model fitting. For each binary genotype pair, the core dataset was divided by stratified random sampling into training (70%), validation (15%) and independent test (15%) subsets (random state = 42). The validation subset was used for model comparison and selection, whereas the independent test subset was reserved for final performance reporting. The source analysis records available for this revision do not document a reproducible exhaustive grid/random search space for the feature-based models; therefore, the previous wording implying formal hyperparameter optimization has been removed rather than reconstructing unrecorded search ranges or selected values retrospectively. Each physical seed contributed one image only, so repeated images of the same seed could not occur across subsets. Because phenotypic-core membership was defined before train/validation/test splitting using genotype-specific centroids, the reported predictive performance applies to the preselected core population and should not be interpreted as an unbiased estimate for all unselected seeds of the source genotypes.

Permutation importance was calculated after model fitting for the best-performing feature-based model in each genotype pair. This analysis was used for post hoc interpretation of predictor contributions and was not used as a pre-classification feature-selection step. Consequently, the reported classification performances reflect models fitted to the full predefined descriptor set rather than to a subset selected from the outcome data.

2.6. Deep learning models

Using the phenotypic core datasets, all possible pairwise combinations among the five flax genotypes were evaluated with deep-learning-based classification models. Each genotype pair was treated as an independent binary classification problem, and all convolutional neural network (CNN) models were compared using the same training and evaluation strategy. The DenseNet121, MobileNetV2, EfficientNetB0 and ResNet50 architectures were used in the analyses [11-14]. These models were selected because they represent different architectural approaches in terms of network depth, number of parameters and feature-extraction strategies, and their performances in learning phenotypic patterns were compared.

CC: BY-NC-ND

All models were developed using a transfer-learning approach. For this purpose, the convolution layers of networks pre-trained on the ImageNet dataset were used as the initial weights; the original classification layer was removed and replaced with a new output layer appropriate for the relevant binary classification problem. The models were then retrained using only the images belonging to the relevant genotype pair. This approach was preferred in order to achieve more effective feature learning from limited-size image datasets, to shorten the training time and to support feature learning with a limited image dataset. All CNN models were trained using the same data-splitting strategy, data-augmentation operations and training training settings, so that the performance differences among the architectures would arise only from the network structure. Adam was used consistently across architectures because it provides adaptive first-order optimization and is widely used for transfer-learning fine-tuning; the study was not designed as an optimizer benchmark, and no claim is made that Adam is superior to SGD, RMSProp or Nadam for these data.

In all CNN models, the input images were rescaled to 224 × 224 pixels and the pixel values were normalized to the 0–1 range. The datasets were divided into training, validation and independent test subsets so that the class ratios were preserved. Data-augmentation operations comprising random horizontal and vertical flipping, 5% rotation, 8% zoom and 10% contrast variation were applied to the training dataset. The models were trained using a two-stage transfer-learning strategy. In the first stage, the pre-trained base network was frozen, and only the new classification layers were trained for 8 epochs using the Adam optimization algorithm and a learning rate of 1×10^{-3}. In the second stage, approximately the last 30% of the base network was unfrozen for training and the model was subjected to fine-tuning for 12 epochs with a learning rate of 1×10^{-5}. In all models, the mini-batch size was set to 16, and the random number generator was fixed at 42 to ensure the reproducibility of the analyses. The nominal 8-epoch frozen-base stage and 12-epoch fine-tuning stage were treated as maximum training windows within the predefined two-stage protocol rather than as exhaustively optimized epoch counts; validation-based callbacks-controlled convergence and could terminate or adjust training when improvement ceased. The phrase 'last 30%' refers to approximately the final 30% of layer objects in each architecture-specific convolutional base, counted from the output end; because the architectures contain different numbers and organizations of layers, the absolute number of trainable layers necessarily differed among models. The percentage was used to apply a comparable relative fine-tuning depth while preserving most pretrained representations.

Validation performance was monitored continuously throughout the training process. To prevent overfitting, the Early Stopping, ReduceLROnPlateau and ModelCheckpoint strategies were applied together. Early Stopping terminated the training process according to the validation AUC value, while ReduceLROnPlateau automatically reduced the learning rate depending on the change in the validation loss. With ModelCheckpoint, the model weights providing the highest AUC value on the validation dataset were saved and used in the final performance evaluations at the end of training.

Model performances were calculated using the evaluation metrics described in Section 2.8. For each binary classification problem, the CNN model providing the highest F1-score was accepted as the best-performing model for the relevant genotype pair.

After the classification performances had been compared, DenseNet121 was used as the common feature extractor for the subsequent feature-space analyses. The deep feature vectors obtained before the final classification layer were used to calculate image-derived phenotypic distances, evaluate similarity relationships among genotypes and construct the common feature space.

2.7. Model interpretation and feature-space analyses

Model interpretation was deliberately restricted to two complementary levels. For feature-based machine-learning models, permutation importance quantified the contribution of measured morphological, colour and texture descriptors. For CNN representations, deep feature-space analyses described genotype separation and within-genotype dispersion. Grad-CAM or other pixel-level class-activation/saliency maps were not generated in the underlying analysis; therefore, the revised manuscript does not claim pixel-level visual explainability or present XAI as a major novelty. This scope is stated explicitly to avoid implying visual evidence that was not produced.

CC: BY-NC-ND

In the first stage, global permutation importance analysis was applied to the feature-based machine-learning models. In this analysis, each morphological, colour, and texture feature was randomly permuted, and its relative importance was determined by calculating the change it produced in the model performance. By evaluating all binary classifications together, the phenotypic features contributing most to genotype discrimination were thus revealed.

To complement the global evaluation, pairwise permutation importance analyses were additionally performed on the genotype pairs showing the highest and the lowest classification success. It was thus examined whether the features driving classification varied among different genotype pairs, and whether the easily and the hardly classified genotype pairs had different phenotypic determinants.

In the deep-learning-based feature-space analyses, the high-dimensional feature vectors obtained before the final classification layer of the DenseNet121 model were used. Using these features, the geometric centroid of each genotype in the multidimensional feature space was calculated as the arithmetic mean of the feature vectors of the relevant samples. The Euclidean distances among the genotype centroids were then determined, and the resulting distance matrix was related to the F1-scores obtained from both the feature-based machine-learning and the CNN models, to evaluate the link between phenotypic separation and discriminability.

To evaluate the relationships among genotypes visually, a phenotypic similarity network was constructed using the distance matrix. In the network structure, the nodes represent the genotypes; node size was scaled according to the mean distance to the centre so as to reflect within-genotype homogeneity, while the edges were weighted according to the highest CNN F1-score obtained for the relevant genotype pair. In addition, hierarchical clustering analysis was applied to the same distance matrix, and the results were visualized as a dendrogram to evaluate the phenotypic organization among genotypes.

To determine within-genotype phenotypic homogeneity, the Euclidean distance of each seed to its own genotype center was calculated. Using these distances, the mean distance to the center, the standard deviation, and the coefficient of variation were determined, and the distributions were visualized using violin and box plots. The degree of dispersion of each genotype in the feature space was thus compared quantitatively.

Finally, the high-dimensional feature space obtained from the DenseNet121 model was reduced to two dimensions using the Uniform Manifold Approximation and Projection (UMAP) method. On the two-dimensional projections, the genotype clusters, the cluster centers, and the 95% confidence ellipses were evaluated together to examine the phenotypic organization visually in the learned feature space and the separation among genotypes.

By evaluating these analyses together, the phenotypic features contributing to classification performance, the distribution of the genotypes in the feature space, the level of within-genotype homogeneity and the relationships between phenotypic similarity and classifiability were assessed within a holistic approach.

2.8. Performance evaluation, statistical analysis and computational environment

Model performance was evaluated using accuracy, precision, recall, the F1-score, the Matthews correlation coefficient (MCC), and the area under the ROC curve (ROC–AUC). The F1-score was adopted as the primary performance metric because it takes precision and recall into account jointly, while the MCC was used as a balanced evaluation metric that is less affected by the class distribution. The ROC curves and ROC–AUC values were calculated to compare the discriminative performances of the models at different decision thresholds. For each binary classification problem, the best-performing model was identified as the model providing the highest F1-score on the independent test dataset.

To determine the relationship between phenotypic distance and classification success, correlation and linear regression analyses were performed. The relationships between the Euclidean distances among genotypes and the F1-scores obtained from the machine-learning and deep-learning models

CC: BY-NC-ND

were evaluated using Pearson and Spearman rank correlation analyses. In addition, linear regression analysis was applied to determine the extent to which phenotypic distance explained classification performance, and the coefficient of determination (R^2) was calculated. The statistical significance of the correlation coefficients was assessed using the corresponding *p-values*.

Differences among genotypes in terms of morphological, color and texture features were evaluated by one-way analysis of variance (ANOVA). For the variables in which a significant difference was determined, Fisher's Least Significant Difference (LSD) multiple comparison test was applied to compare the means. For within-genotype dispersion, a one-way ANOVA was performed using the individual seed-to-centroid distances in the DenseNet121 feature space, and the homogeneity-of-variance assumption was evaluated using the Brown–Forsythe test. In all statistical analyses, the significance level was accepted as $P < 0.05$, and the results are presented as mean ± standard deviation where applicable.

All of the data processing, image analysis, machine-learning, deep-learning and statistical analyses of the study were carried out using the Python 3.12 programming language. NumPy and Pandas were used for data management and numerical computation, SciPy and Statsmodels for the statistical analyses, Scikit-learn for the machine-learning applications, TensorFlow and Keras for the deep-learning models, OpenCV for the image-processing operations, NetworkX for the network analyses, UMAP-learn for the dimensionality-reduction analyses and Matplotlib for preparing the figures. Throughout the analyses, the random number generators were fixed (*random seed* = 42), and the reproducibility of all computations was ensured.

2.9. Workflow of the analytical framework

The overall analytical workflow is summarized in Figure 1. It comprises extraction of morphological, color and texture features from standardized single-seed images, construction of phenotypic core datasets, pairwise feature-based and deep-learning classification, permutation-importance interpretation, and deep feature-space analyses. This integrated framework was used to evaluate phenotypic similarity, within-genotype dispersion, and classification performance on a common database.

3. Results

Statistically significant differences were determined among the five flax genotypes in terms of morphological, color and texture features ($P < 0.001$; Table 1). The morphological evaluations showed that the EBYC01 genotype generally had a larger seed morphology, whereas K1104 had a smaller one. This difference was particularly pronounced in the seed area (6.21 mm^2), width (4.66 mm), major axis length (4.64 mm), and equivalent diameter (2.80 mm) variables. In contrast, the Culbert, EBYC01, and K0902 genotypes exhibited similar means in terms of aspect ratio, while BeyazGelin and K1104 showed lower values. Significant differences among genotypes were also determined in the circularity and solidity variables. Among the morphological features, the highest F values were calculated for width ($F = 157.38$) and major axis length ($F = 142.77$) (Table 1).

Similar to the differences observed in the morphological features, the color features also revealed marked variation among genotypes (Table 1). The EBYC01 genotype had the highest mean values for the red, green, blue, value, and lightness components, while Culbert exhibited the lowest values, particularly for the green and blue components. The hue variable was observed at the highest level in the BeyazGelin genotype and at the lowest levels in the K0902 and K1104 genotypes. Significant differences were also detected for the a* and b* coordinates of the CIELAB color space. Among the color variables, the lowest coefficients of variation were determined for the a* (0.84%) and correlation (1.15%) variables (Table 1).

Marked differences among genotypes were also detected in terms of texture features (Table 1). Significant differences were determined for all the contrast, dissimilarity, homogeneity, energy, correlation, ASM and entropy variables ($P < 0.001$). The EBYC01 genotype had the highest mean values for contrast and entropy, while BeyazGelin showed higher means for the homogeneity and energy

CC: BY-NC-ND

variables. The ASM variable showed the highest coefficient of variation (26.41%) among all texture features. Fisher's LSD test confirmed that there were significant differences among genotypes for all texture features (Table 1).

For the five flax genotypes, all 10 unique genotype pairs were evaluated as independent pairwise binary classification tasks using the feature-based machine-learning models. The F1-scores ranged from 0.79 to 0.97 (Table 2; Figure 2). The highest classification success was obtained for the BeyazGelin–Culbert pair with the SVM (RBF) model; an accuracy of 96.7%, an F1-score of 0.967 and an MCC of 0.933 were determined. Similarly, high performance values were obtained with the MLP and Random Forest models for the Culbert–EBYC01 and EBYC01–K1104 pairs, respectively. In contrast, the lowest performance was observed for the Culbert–K0902 pair with the Random Forest model, where an F1-score of 0.792 and an MCC of 0.651 were calculated. When all 10 pairwise binary classifications were evaluated together, the mean accuracy was 90%, and the mean F1-score was 0.90 (Table 2).

The global permutation importance analysis showed that the Hue and a* variables had the highest mean importance values (Figure 3). These were followed by the width and major axis length variables, while among the texture features, correlation, dissimilarity, and contrast exhibited the highest importance values. When the binary classifications with the highest and the lowest F1-scores were compared, the relative importance rankings of the variables were observed to differ. In the BeyazGelin–Culbert pair, the morphological and color features showed higher importance values, whereas in the Culbert–K0902 pair, the Hue and dissimilarity variables came to the fore (Figure 4).

Table 1. Mean ± standard deviation of morphological, color, and texture traits measured for five flax genotypes. Different lowercase letters within a row indicate significant differences among genotypes according to Fisher's LSD test ($P < 0.05$).

Trait	BeyazGelin	Culbert	EBYC01	K0902	K1104	F value	P value	CV (%)
Morphological traits								
Area (mm^2)	5.43 ± 0.99 b	5.24 ± 0.86 c	6.21 ± 0.91 a	5.23 ± 0.78 c	5.06 ± 0.69 d	56.47	<0.001	15.68
Perimeter (mm)	12.26 ± 1.27 a	11.46 ± 0.82 b	12.43 ± 1.01 a	11.21 ± 0.79 c	11.10 ± 0.78 c	82.71	<0.001	8.16
Width (mm)	4.48 ± 0.29 b	4.27 ± 0.22 c	4.66 ± 0.26 a	4.22 ± 0.21 d	4.14 ± 0.21 e	157.38	<0.001	5.49
Height (mm)	2.07 ± 0.26 a	1.91 ± 0.23 b	2.10 ± 0.24 a	1.91 ± 0.23 b	1.94 ± 0.17 b	35.53	<0.001	11.37
Major axis length (mm)	4.51 ± 0.36 b	4.16 ± 0.25 c	4.64 ± 0.31 a	4.12 ± 0.25 c	4.14 ± 0.25 c	142.77	<0.001	6.66
Minor axis length (mm)	1.78 ± 0.28 b	1.71 ± 0.23 c	1.83 ± 0.23 a	1.72 ± 0.21 c	1.69 ± 0.19 c	13.48	<0.001	13.09
Equivalent diameter (mm)	2.62 ± 0.24 b	2.58 ± 0.21 c	2.80 ± 0.21 a	2.57 ± 0.19 c	2.53 ± 0.17 d	54.26	<0.001	7.85
Aspect ratio	2.20 ± 0.32 b	2.27 ± 0.28 a	2.24 ± 0.29 a	2.24 ± 0.27 a	2.15 ± 0.21 b	5.82	<0.001	12.55
Circularity	0.46 ± 0.09 c	0.50 ± 0.08 b	0.51 ± 0.07 b	0.52 ± 0.06 a	0.52 ± 0.07 a	22.93	<0.001	14.70
Extent	0.58 ± 0.07 c	0.64 ± 0.05 a	0.64 ± 0.06 b	0.65 ± 0.04 a	0.63 ± 0.06 b	43.84	<0.001	8.76
Solidity	0.85 ± 0.08 c	0.90 ± 0.05 b	0.91 ± 0.04 a	0.92 ± 0.03 a	0.90 ± 0.05 b	46.73	<0.001	5.82
Color traits								
Red	139.44 ± 6.34 d	138.22 ± 5.30 e	146.03 ± 5.41 a	142.28 ± 4.58 b	140.98 ± 5.14 c	62.62	<0.001	3.81
Green	90.89 ± 8.21 b	85.53 ± 5.98 c	97.43 ± 6.20 a	90.02 ± 5.45 b	90.45 ± 5.72 b	88.70	<0.001	7.03
Blue	75.65 ± 9.00 b	66.57 ± 6.71 d	78.61 ± 6.54 a	71.02 ± 6.18 c	74.50 ± 6.18 b	87.37	<0.001	9.56
Hue	11.78 ± 3.58 a	10.85 ± 1.68 b	10.02 ± 1.49 c	9.16 ± 1.12 d	9.36 ± 2.25 d	48.81	<0.001	21.49
Saturation	121.71 ± 11.87 d	137.48 ± 9.48 a	122.50 ± 8.41 d	132.38 ± 8.56 b	124.83 ± 8.42 c	105.80	<0.001	7.39
Value	139.45 ± 6.34 d	138.22 ± 5.30 e	146.03 ± 5.41 a	142.28 ± 4.58 b	140.98 ± 5.14 c	62.59	<0.001	3.81
Lightness	111.00 ± 7.62 b	106.82 ± 5.68 c	117.34 ± 5.84 a	111.18 ± 5.08 b	111.15 ± 5.44 b	78.63	<0.001	5.38
a*	146.52 ± 1.49 d	148.20 ± 1.17 a	145.85 ± 1.20 e	147.74 ± 1.17 b	147.24 ± 1.12 c	115.90	<0.001	0.84
b*	145.48 ± 2.11 d	148.72 ± 2.05 a	147.20 ± 1.78 b	148.38 ± 1.78 a	146.24 ± 1.77 c	104.64	<0.001	1.29
Texture traits								
Contrast	5.71 ± 0.81 d	6.55 ± 0.80 b	6.82 ± 0.96 a	6.30 ± 0.89 c	5.37 ± 0.65 e	105.46	<0.001	13.47
Dissimilarity	0.93 ± 0.08 c	1.06 ± 0.09 a	1.06 ± 0.10 a	1.00 ± 0.08 b	0.92 ± 0.08 c	135.02	<0.001	8.69
Homogeneity	0.76 ± 0.02 a	0.73 ± 0.02 d	0.73 ± 0.03 d	0.74 ± 0.02 c	0.75 ± 0.02 b	97.41	<0.001	2.82
Energy	0.39 ± 0.06 a	0.34 ± 0.04 b	0.35 ± 0.05 b	0.34 ± 0.03 c	0.35 ± 0.05 b	49.52	<0.001	13.07
Correlation	0.93 ± 0.01 a	0.92 ± 0.01 d	0.93 ± 0.01 b	0.92 ± 0.01 c	0.94 ± 0.01 a	110.86	<0.001	1.15
ASM	0.16 ± 0.05 a	0.12 ± 0.03 c	0.12 ± 0.03 b	0.11 ± 0.02 c	0.13 ± 0.04 b	54.40	<0.001	26.41
Entropy	55.99 ± 3.41 d	59.98 ± 2.48 b	60.91 ± 3.06 a	59.96 ± 2.17 b	58.18 ± 3.05 c	92.46	<0.001	4.86

Values are mean ± SD. Different lowercase letters within a row indicate significant differences among genotypes according to Fisher's LSD test ($P < 0.05$). Area is expressed as mm^2, linear morphological traits as mm, and colour variables are reported in the original 8-bit image scale.

CC: BY-NC-ND

Table 2. Classification performance of the best feature-based machine learning model for each binary flax genotype pair. Performance was evaluated using the independent test dataset. The best-performing model for each genotype pair is reported based on the highest F1-score.

Pair	Genotype 1	Genotype 2	Best ML model	Accuracy	Precision	Recall	F1 score	MCC	ROC-AUC
BeyazGelin vs. Culbert	BeyazGelin	Culbert	SVM (RBF)	0.97	0.97	0.97	0.97	0.93	0.98
Culbert vs. EBYC01	Culbert	EBYC01	MLP	0.95	0.97	0.93	0.95	0.90	0.99
EBYC01 vs. K1104	EBYC01	K1104	Random Forest	0.93	0.91	0.97	0.94	0.87	0.98
BeyazGelin vs. EBYC01	BeyazGelin	EBYC01	SVM (RBF)	0.93	0.96	0.90	0.93	0.87	0.99
Culbert vs. K1104	Culbert	K1104	Random Forest	0.92	0.93	0.90	0.92	0.83	0.96
BeyazGelin vs. K0902	BeyazGelin	K0902	Random Forest	0.92	0.93	0.90	0.92	0.83	0.97
EBYC01 vs. K0902	EBYC01	K0902	XGBoost	0.92	0.93	0.90	0.92	0.83	0.97
BeyazGelin vs. K1104	BeyazGelin	K1104	SVM (RBF)	0.85	0.84	0.87	0.85	0.70	0.90
K0902 vs. K1104	K0902	K1104	XGBoost	0.83	0.81	0.87	0.84	0.67	0.93
Culbert vs. K0902	Culbert	K0902	Random Forest	0.82	0.91	0.70	0.79	0.65	0.89

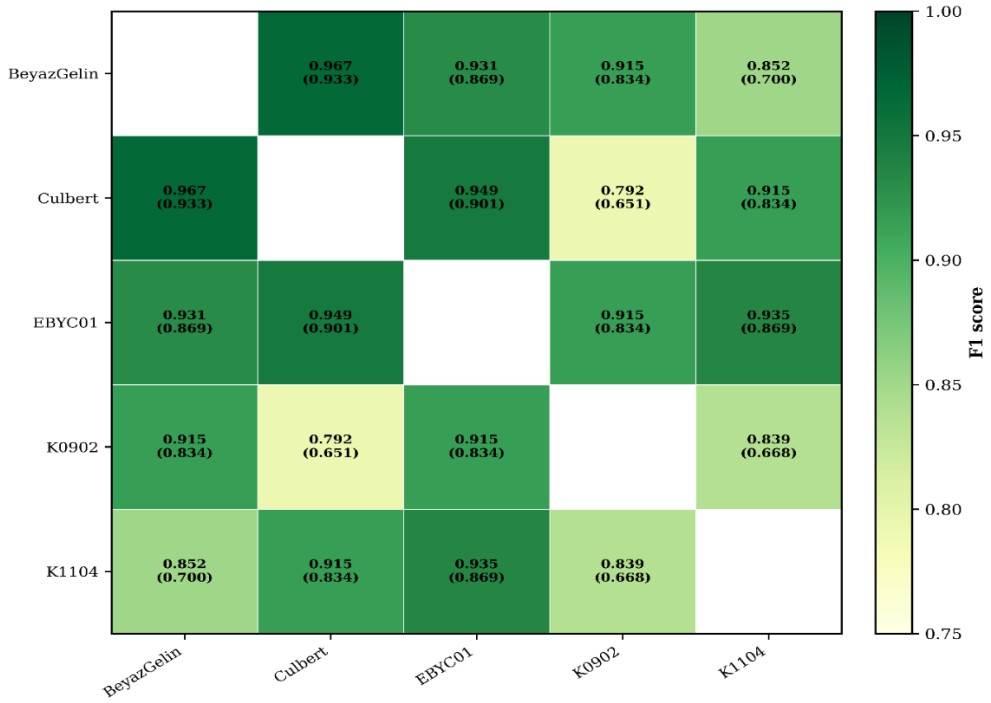

Figure 2. Pairwise binary classification performance of feature-based machine-learning models for the five flax genotypes. The five genotypes generated 10 unique pairwise binary classification tasks. Each off-diagonal cell represents one genotype pair and shows the highest F1-score obtained among the evaluated feature-based machine-learning models, with the corresponding Matthews correlation coefficient (MCC) given in parentheses. The matrix is displayed symmetrically for visualization, and diagonal cells are left blank because no within-genotype classification was performed. Darker green shading indicates higher F1-score.

CC: BY-NC-ND

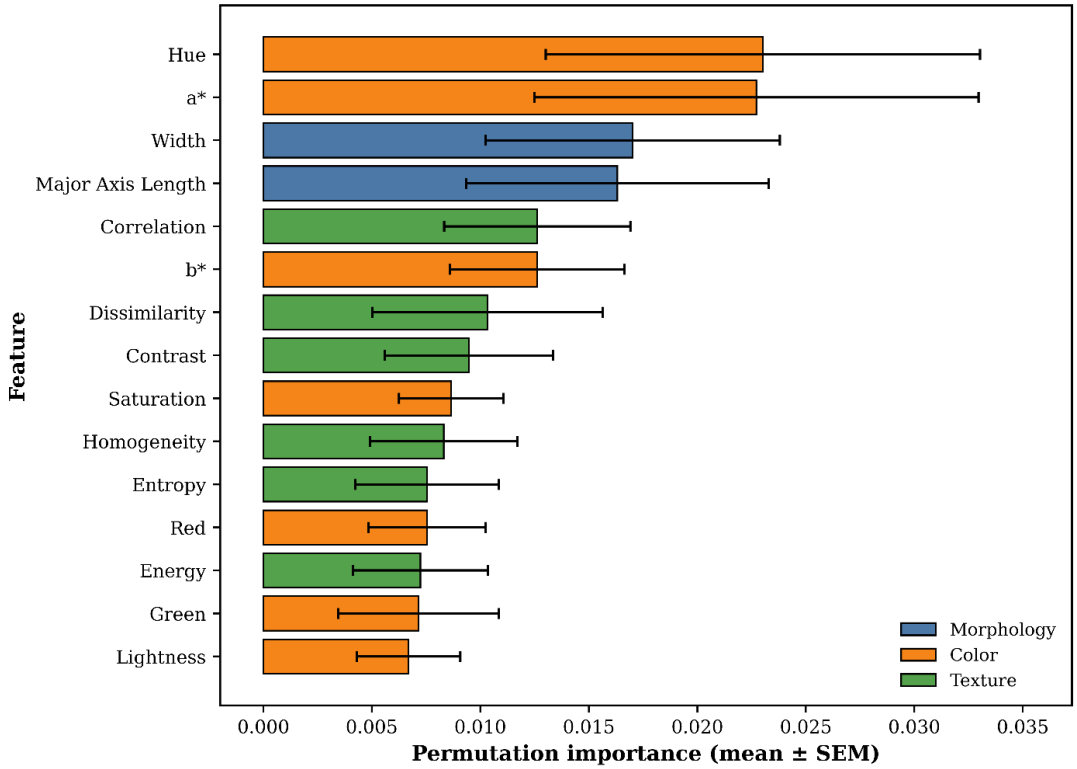

Figure 3. Global permutation importance of morphological, colour, and texture features across the best-performing models for the 10 pairwise binary genotype-classification tasks. Bars represent the mean permutation importance of each feature across the evaluated pairwise tasks, and error bars represent the standard error of the mean (SEM). Features are grouped by descriptor category as morphological, colour, or texture traits.

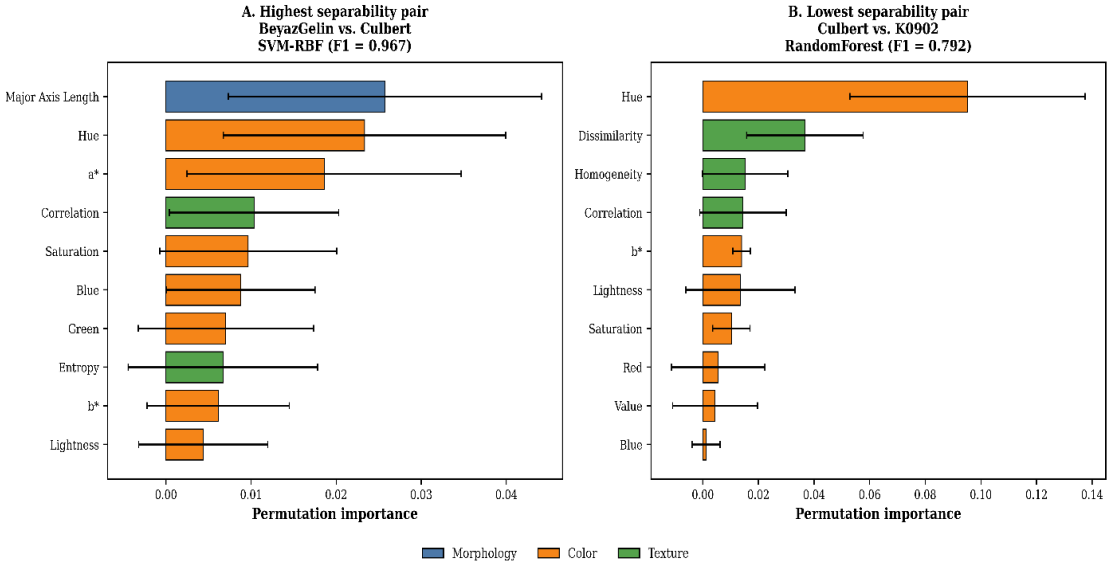

Figure 4. Comparison of feature importance between the easiest and most difficult pairwise genotype-classification tasks. Permutation importance values are shown for (A) the highest-performing genotype pair, BeyazGelin–Culbert, classified using SVM-RBF (F1 = 0.967), and (B) the lowest-performing pair, Culbert–K0902, classified using Random Forest (F1 = 0.792). Bars represent mean permutation importance, error bars represent the standard error of the mean (SEM), and colours distinguish morphological, colour, and texture traits.

CC: BY-NC-ND

In the binary classifications performed with the convolutional neural network models, the F1-scores ranged from 0.706 to 0.951 (Table 3). The highest performance was obtained for the K0902–K1104 pair using the MobileNetV2 model, with an accuracy of 95.0%, an F1-score of 0.951, and an MCC of 0.901. The DenseNet121 model provided the highest F1-score in five of the ten binary classifications. With this model, F1-scores of 0.923, 0.923 and 0.909 were obtained for the BeyazGelin–Culbert, BeyazGelin–K0902 and Culbert–K1104 pairs, respectively. The lowest performance was observed for the EBYC01–K0902 pair, with an F1-score of 0.706 (Table 3).

Table 3. Performance of the best convolutional neural network model for each binary flax genotype classification task.

Pair	Genotype 1	Genotype 2	DenseNet121 feature-space distance (Euclidean)	Best CNN model	Accuracy	Precision	Recall	F1 score	MCC	ROC-AUC
K0902 vs. K1104	K0902	K1104	12.46	MobileNetV2	0.95	0.94	0.97	0.95	0.90	0.98
BeyazGelin vs. Culbert	BeyazGelin	Culbert	13.34	DenseNet121	0.92	0.86	1.00	0.92	0.85	0.98
BeyazGelin vs. K0902	BeyazGelin	K0902	12.69	DenseNet121	0.92	0.86	1.00	0.92	0.85	0.98
Culbert vs. K1104	Culbert	K1104	14.67	DenseNet121	0.92	1.00	0.83	0.91	0.85	1.00
Culbert vs. K0902	Culbert	K0902	10.13	DenseNet121	0.85	0.80	0.93	0.86	0.71	0.98
BeyazGelin vs. EBYC01	BeyazGelin	EBYC01	11.11	MobileNetV2	0.87	0.96	0.77	0.85	0.75	0.94
EBYC01 vs. K1104	EBYC01	K1104	13.01	MobileNetV2	0.77	0.68	1.00	0.81	0.60	0.99
Culbert vs. EBYC01	Culbert	EBYC01	10.06	DenseNet121	0.83	0.96	0.70	0.81	0.69	0.93
BeyazGelin vs. K1104	BeyazGelin	K1104	9.36	MobileNetV2	0.75	0.67	1.00	0.80	0.58	0.96
EBYC01 vs. K0902	EBYC01	K0902	9.44	DenseNet121	0.58	0.55	1.00	0.71	0.30	0.94

Performance metrics were calculated using the independent test dataset. For each genotype pair, only the convolutional neural network model achieving the highest F1-score is presented. Distances represent Euclidean distances between genotype centroids extracted from the DenseNet121 feature space

The phenotypic similarity network constructed from the DenseNet121 feature space visually revealed the relationships among genotypes (Figure 5). In the network structure, node size represented within-genotype homogeneity, while edge thickness and colour intensity represented the highest CNN F1-score obtained for the relevant genotype pair. In the network, BeyazGelin and K1104 were represented by larger nodes, whereas Culbert and EBYC01 were represented by smaller nodes.

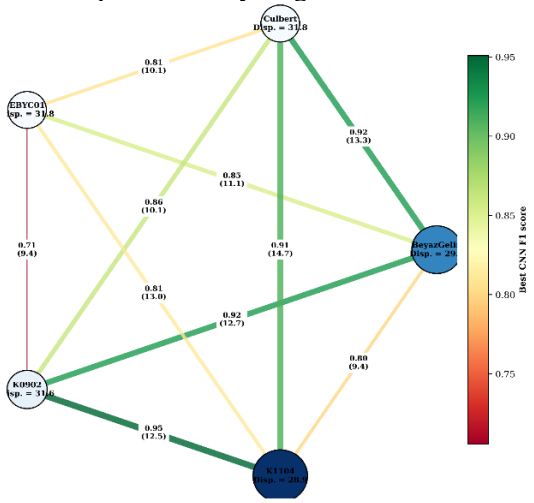

Figure 5. Phenotypic similarity network constructed from DenseNet121 feature-space distances. Node size is proportional to within-genotype homogeneity, whereas edge color and thickness represent the highest CNN F1-score achieved for each genotype pair. Numbers in parentheses indicate the corresponding feature-space distances.

CC: BY-NC-ND

The pairwise Euclidean distances calculated from the same feature space ranged from 9.36 to 14.67 (Table 4). The smallest distances were calculated for the BeyazGelin–K1104 (9.36) and EBYC01–K0902 (9.44) pairs, while the Culbert–K1104 pair showed the highest distance value (14.67). The hierarchical clustering analysis based on these distances is presented in Figure 6. Consistent with the distance matrix, the dendrogram showed that BeyazGelin and K1104 were in the same main cluster, while EBYC01 and K0902 were in the second main cluster.

Table 4. Pairwise Euclidean distances between flax genotype centroids in the DenseNet121 feature space.

Genotype	BeyazGelin	Culbert	EBYC01	K0902	K1104
BeyazGelin	0.00	13.34	11.11	12.69	9.36
Culbert	13.34	0.00	10.06	10.13	14.67
EBYC01	11.11	10.06	0.00	9.44	13.01
K0902	12.69	10.13	9.44	0.00	12.46
K1104	9.36	14.67	13.01	12.46	0.00

Distances were calculated as Euclidean distances between genotype centroids in the standardized DenseNet121 deep feature space. Larger values indicate greater phenotypic dissimilarity.

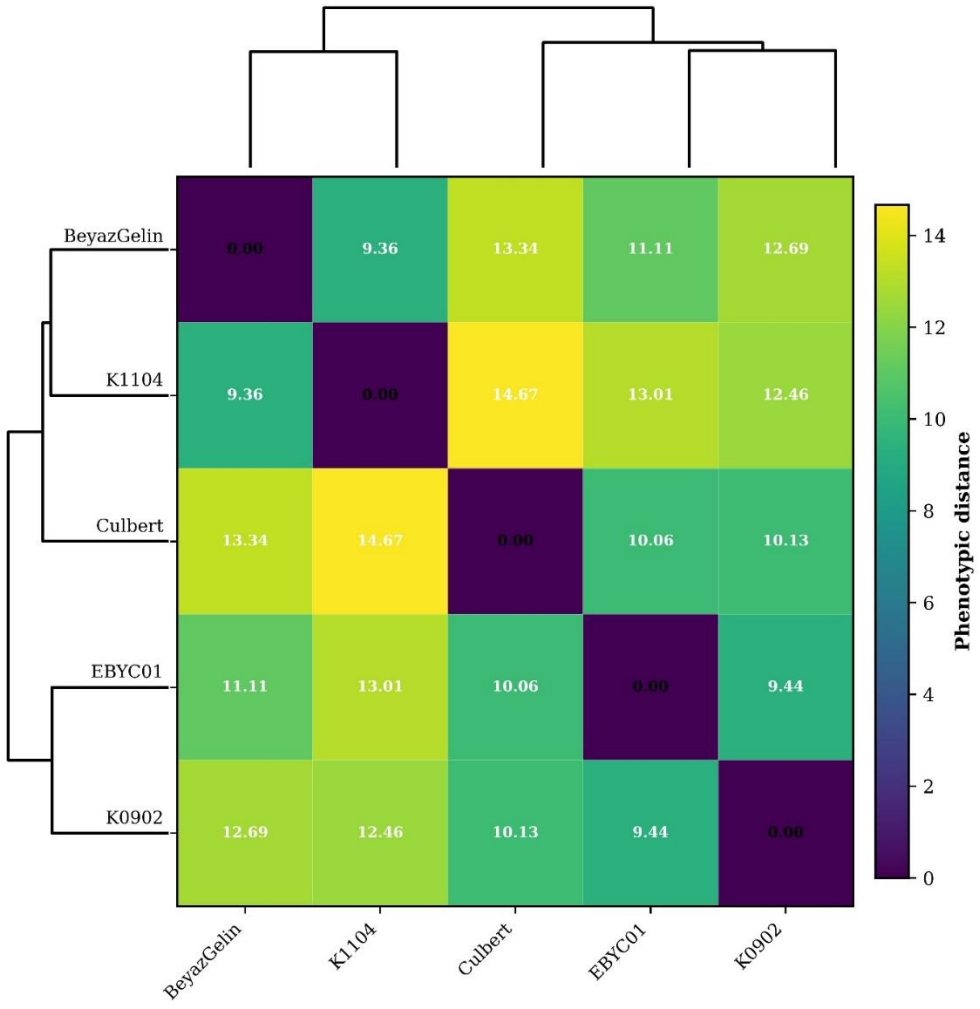

Figure 6. Hierarchically clustered pairwise Euclidean distance matrix of flax genotypes in the DenseNet121 feature space. Smaller distances indicate greater phenotypic similarity, whereas larger distances reflect increased separation between genotype centroids.

As a result of the within-genotype dispersion analysis, the mean distance to the genotype-specific centroid ranged from 28.890 to 31.796 (Table 5; Figure 7). The lowest mean dispersion was observed for K1104 (28.890) and BeyazGelin (29.827), whereas EBYC01 (31.796), Culbert (31.773), and K0902 (31.565) showed higher mean values. In contrast, the coefficients of variation were relatively similar across genotypes, ranging from 13.661% to 15.677%. A one-way ANOVA based on the individual seed-to-centroid distances revealed significant differences in within-genotype dispersion among the five genotypes ($F(4, 995) = 17.287$, $P < 0.001$). The homogeneity-of-variance assumption was supported by the Brown–Forsythe test ($P = 0.622$).

Table 5. A one-way ANOVA based on the individual seed-to-centroid distances revealed significant differences in within-genotype dispersion among the five genotypes (F (4,995) = 17.287, P < 0.001).

Genotype	N	Mean dispersion	SD	SEM	Median	Minimum	Maximum	CV (%)
BeyazGelin	200	29.827	4.66	0.33	28.817	22.389	48.907	15.623
Culbert	200	31.773	4.638	0.328	31.369	22.271	57.934	14.598
EBYC01	200	31.796	4.344	0.307	31.233	24.036	50.301	13.661
K0902	200	31.565	4.503	0.318	31.516	21.804	51.794	14.267
K1104	200	28.89	4.529	0.32	27.763	22.764	48.727	15.677

Dispersion was calculated as the Euclidean distance of each individual seed from the centroid of its corresponding genotype in the 1,024-dimensional DenseNet121 feature space. A one-way ANOVA based on the individual seed-level distances indicated significant differences in within-genotype dispersion among genotypes [F(4, 995) = 17.287, P < 0.001]. The homogeneity-of-variance assumption was supported by the Brown–Forsythe test (P = 0.622). SD, standard deviation; SEM, standard error of the mean; CV, coefficient of variation.

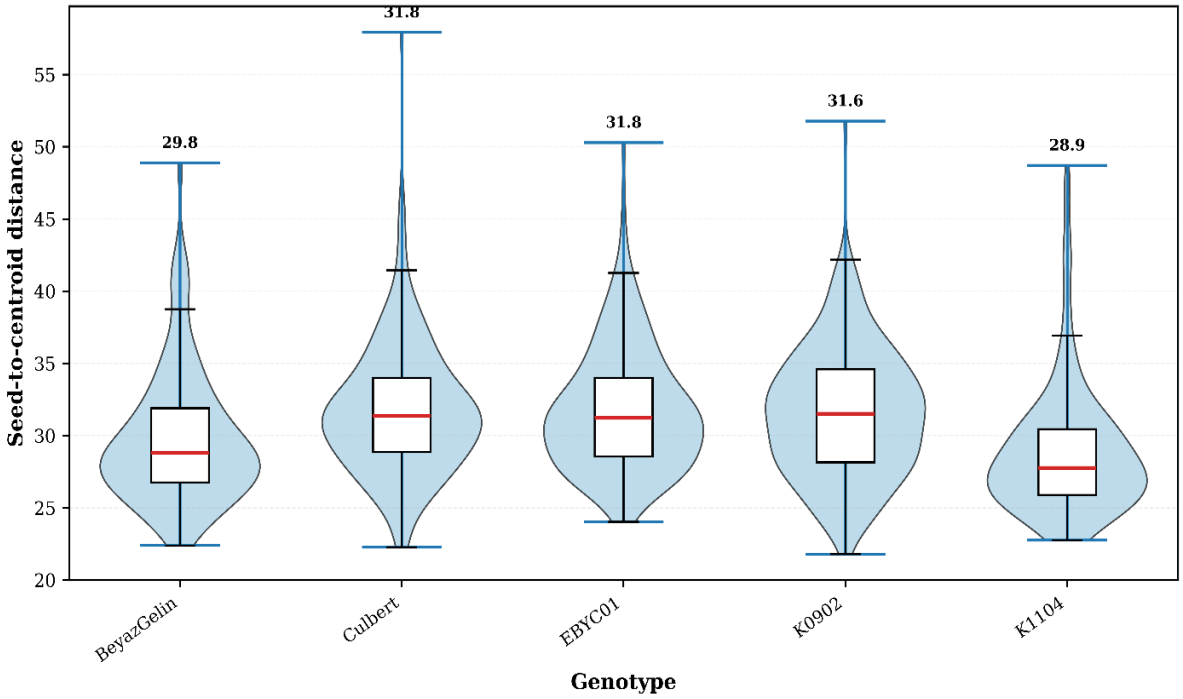

Figure 7. Within-genotype dispersion in the DenseNet121 feature space. Violin plots combined with boxplots illustrate the distribution of seed-to-centroid distances for each flax genotype, providing a measure of intra-genotype phenotypic variability.

The relationship between DenseNet121 feature-space distance and classification success was evaluated using Pearson and Spearman correlation analyses (Table 6; Figure 8). In the feature-based machine-learning models, a positive but not statistically significant relationship was determined between feature-space distance and the F1-score (Pearson r = 0.346, P = 0.328). In contrast, for the convolutional neural network models, the same relationship was found to be stronger and statistically significant (Pearson r = 0.710, P = 0.021). The Spearman rank correlation showed a similar trend (ρ = 0.699, P = 0.025), and the coefficient of determination in the linear regression analysis was calculated as 50.4% (R^2 = 0.504). In addition, the relationship between feature-space distance and classification success was found to be stronger in the CNN models than in the conventional machine-learning models.

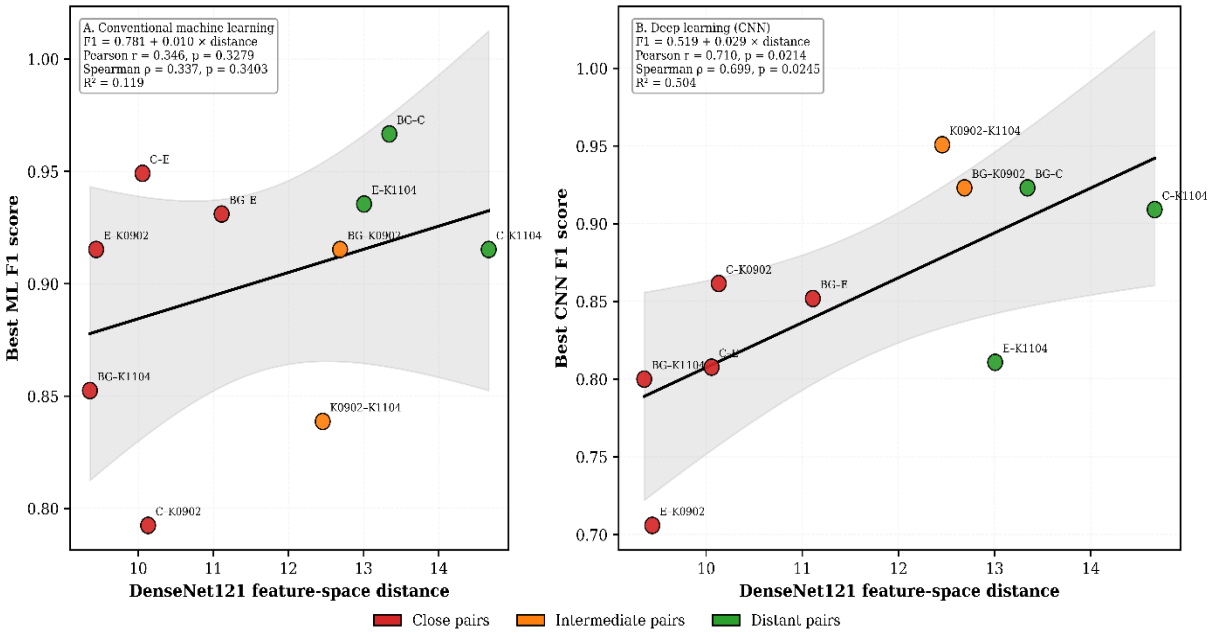

Figure 8. Relationship between DenseNet121 feature-space distance and binary classification performance. Linear regression lines with 95% confidence intervals illustrate the association between feature-space distance and F1-score for feature-based machine learning models (A) and convolutional neural network models (B).

Table 6. Correlation between DenseNet121 feature-space distance and binary classification performance.

Model family	Feature-based machine learning	Convolutional neural network
Response variable	ML F1 score	CNN F1 score
N pairs	10	10
Regression equation	F1 = 0.781 + 0.010 × distance	F1 = 0.519 + 0.029 × distance
Intercept	0.7814	0.5189
Slope	0.0103	0.0289
Pearson r	0.3457	0.7102
Pearson P value	0.3279	0.0214
Spearman rho	0.3374	0.6991
Spearman P value	0.3403	0.0245
R^2	0.1195	0.5043
Significance	ns	*

CC: BY-NC-ND

The two-dimensional UMAP projection obtained from the DenseNet121 deep features visualised the overall distribution of the samples of the five genotypes in the learned feature space (Figure 9). Although there were partial overlaps among the genotype clusters, each genotype was clustered around its own distinctive centre, and the cluster centres were located at different positions in the feature space.

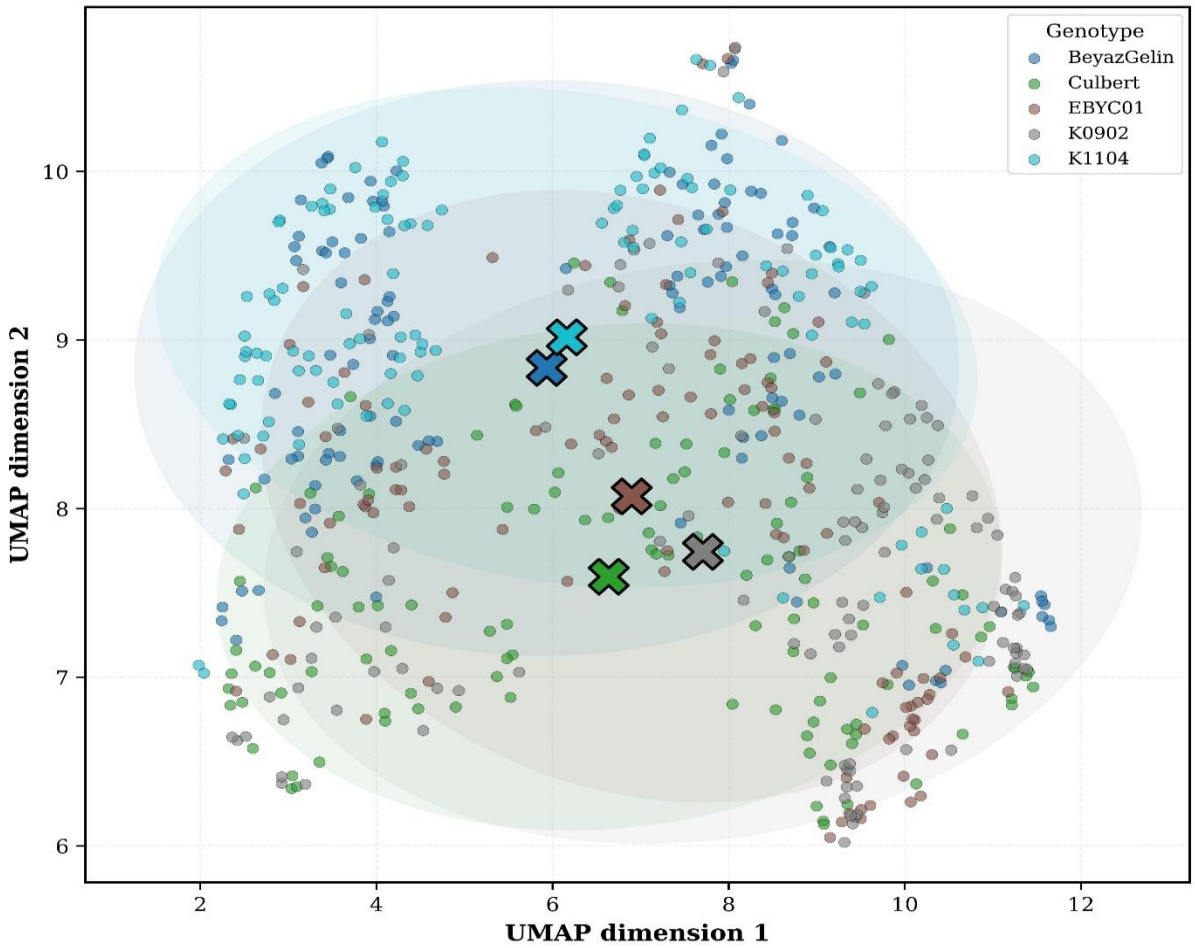

Figure 9. UMAP projection of DenseNet121 deep feature representations for five flax genotypes. Points correspond to individual seeds, crosses indicate genotype centroids, and shaded ellipses represent the overall spatial distribution of each genotype in the learned feature space.

Pairwise discriminability depended on both the genotype pair and the representation/model family. Feature-based machine learning identified BeyazGelin-Culbert as the easiest pair (F1 = 0.967) and Culbert-K0902 as the most difficult (F1 = 0.792), whereas the CNN analysis produced a different ranking for some pairs. For example, Culbert-K0902 reached a higher best-CNN F1 score (0.857) than in the feature-based analysis. This divergence is expected because the two model families operate on different representations: the classical models use predefined morphological, color, and GLCM descriptors, whereas CNNs learn task-specific image representations. Accordingly, disagreement between ML and CNN rankings should not be interpreted as a contradiction in the biological data, but as evidence that pairwise separability is representation-dependent.

4. Discussion

4.1. Binary classification reveals pair-dependent discriminability among flax genotypes

The binary classification analyses performed in this study revealed that the level of distinguishability among flax genotypes varied considerably depending on the genotype pair. Although

the four different CNN architectures generally showed high classification performance, marked performance differences among genotype pairs were observed (Figure 1 and Figure 2). In particular, while the K0902–K1104 comparison achieved the highest reported CNN F1 score (0.951), the performance of other pairs varied, showing that the phenotypic separation among genotypes is not homogeneous. This indicates that image-based classification success depends not only on the model architecture used but also on the phenotypic characteristics of the compared genotypes.

Although the phenotypic information obtained from seed images is not a direct reflection of the genetic make-up, it provides important indicators that make the differences among genotypes visible. Traits such as seed shape, size, color distribution and surface texture develop under genetic control and, although they are influenced to a certain extent by environmental effects, they can create distinctive patterns among genotypes [2]. Therefore, the fact that some genotype pairs can be separated with high accuracy by the deep-learning models suggests that these genotypes carry more pronounced phenotypic differences. In contrast, the comparisons showing lower success indicate that the relevant genotypes may have phenotypic structures that are closer to one another in terms of seed morphology and surface characteristics.

Similar results have been reported in image-based classification studies of flax. Wiesnerová and Wiesner [15] examined seed shape and color for cultivar description, while Smykalova et al. [16] evaluated flax seed phenotypes using image analysis. Recent advances in flax breeding and genome assembly have expanded knowledge of flax genetic resources and their genomic characteristics [17, 18]. Image-based seed phenotyping offers a complementary, non-destructive approach for characterizing observable variation among genotypes [15, 16]. Unlike multiclass accuracy alone, the present study also examined pairwise discrimination and its association with learned feature-space relationships.

The results obtained show that deep-learning models should not be regarded merely as classification tools providing high accuracy in agricultural image analysis. CNN models can learn the complex patterns in images by simultaneously evaluating a large number of phenotypic features such as shape, size, color distribution and surface texture, and they can thereby make visible phenotypic relationships that classical univariate measurements cannot reveal. In particular, the performance differences between the most easily and the most difficultly separated genotype pairs suggest that some genotypes are separated more clearly in terms of morphological and visual characteristics, while others exhibit phenotypically closer characteristics. This shows that classification success depends not only on the CNN architecture used but also on the level of phenotypic separation among genotypes. Indeed, the phenotypic distance, within-genotype homogeneity and feature-space analyses presented in the following sections also support this interpretation, revealing that the differences in classification performance are closely related to the phenotypic organization in the learned deep feature space.

Another notable result is that the overall performance rankings of the different CNN architectures were largely similar. Although there were small performance differences among the models, the most easily and the most difficultly separated genotype pairs were identified similarly in most architectures. This suggests that the observed performance differences are related more to the biological structure of the dataset than to the model. In other words, the high-accuracy separation of some genotype pairs, or the increased difficulty of separation, arises largely from the phenotypic characteristics carried by these genotypes. This result shows that image-based deep-learning approaches can be used not only for prediction but also for understanding and interpreting the phenotypic relationships among genotypes.

Overall, the binary classification results show that there are considerable phenotypic differences among flax genotypes and that these differences can be successfully captured by deep-learning models. At the same time, the performance variability among genotype pairs reveals that the concept of classifiability is not merely an algorithmic outcome but also a reflection of the biological and phenotypic relationships among genotypes. Therefore, binary classification performance can be evaluated not only as a metric showing model success but also as a quantitative indicator of the phenotypic similarity and separation among genotypes. This approach shows that image-based

CC: BY-NC-ND

phenotyping has important potential not only for automatic identification but also for the characterization of genetic resources and the biological interpretation of phenotypic relationships.

4.2. Color, morphology and texture jointly drive genotype discrimination

The feature-importance analyses showed that color, morphology and texture features jointly play a role in distinguishing flax genotypes (Figure 3 and Figure 4). However, not all features contributed equally to classification performance, and some variables were observed to be markedly more effective in genotype discrimination. In particular, among the color components, the Hue, a* and b* values; among the morphological features, Width and Major Axis Length; and among the textural features, the Contrast and Entropy variables came to the fore in many pairwise comparisons. This shows that the phenotypic differentiation of flax seeds cannot be explained by a single feature group and that genotype discrimination is shaped by a multidimensional phenotypic structure.

Seed-coat color was an informative descriptor in the feature-based analyses. Hue and the encoded color descriptors contributed to discrimination among some genotype pairs, although their importance varied by comparison. Previous flax seed image-analysis studies have shown that seed color and shape are useful for cultivar characterization [15, 16]. The present feature-based machine-learning results are consistent with that observation. However, permutation importance of handcrafted descriptors does not directly explain which image regions or features the CNNs used to make predictions.

The morphological features also made important contributions to genotype discrimination. In particular, the fact that the Width and Major Axis Length variables had high importance scores for many genotype pairs shows that seed size and overall geometry are among the discriminative characteristics. Seed size and shape are complex characters that are largely under genetic control and, although they can be influenced to a certain extent by environmental conditions, they can exhibit stable phenotypic differences among genotypes [2]. It is therefore understood that the CNN models incorporated into their decision-making process not only color differences but also subtle changes in seed geometry. In particular, the increased importance of morphological characters in distinguishing genotypes with similar color characteristics shows that the different phenotypic information layers work in a complementary manner. This finding reveals that seed morphology, rather than being a sufficient discriminative characteristic on its own, makes an important contribution to genotype discrimination when evaluated together with color and texture characteristics.

Texture descriptors, including Contrast and Entropy, contributed to discrimination in several feature-based pairwise models. These descriptors quantify spatial variation in image intensity and can complement morphological and color measurements. Their importance in the feature-based models should not be interpreted as evidence that the CNNs attended to particular seed-coat regions; pixel-level CNN attribution was not evaluated in this study.

The relative importance of color, morphology, and texture descriptors varied across genotype pairs in the feature-based analyses. Thus, no single handcrafted descriptor group was uniformly dominant. CNN models learned image representations independently of these predefined descriptors; the current analyses do not establish a direct mapping between handcrafted feature importance and CNN decision mechanisms. The observed pairwise performance differences should therefore be interpreted as representation-dependent patterns rather than evidence of a specific visual mechanism.

Overall, in the discrimination of flax genotypes, color, morphology and texture features served as complementary sources of information. While color features generally formed the first level of discrimination, the morphological and textural features contributed to revealing finer phenotypic differences. This finding shows that focusing on a single feature group may not be sufficient for genotype identification and that multidimensional phenotypic information must be evaluated together for high accuracy. In addition, the patterns revealed by the feature-importance analyses support the biological basis of the phenotypic distance, within-genotype homogeneity and feature-space analyses to be discussed in the following sections. Image-based phenotyping can therefore be regarded not merely as an approach providing automatic classification but as a holistic analytical framework that

CC: BY-NC-ND

contributes to revealing the phenotypic organization and the discriminative characteristics among genotypes.

4.3. Phenotypic distance is strongly associated with CNN classification performance

A significant positive association was observed between DenseNet121 centroid distance and best CNN F1 score across the ten genotype pairs (Pearson r = 0.710, P = 0.021; Spearman rho = 0.699, P = 0.025; R2 = 0.504). This relationship is moderate-to-strong at the pair level but is not deterministic. Individual exceptions are therefore expected. For instance, K0902-K1104 achieved the highest CNN F1 score (0.951) despite a feature-space distance of 12.46, whereas BeyazGelin-Culbert had a larger distance (13.34) but a slightly lower CNN F1 score (0.923). Such deviations indicate that centroid distance captures only one component of classifiability; within-genotype dispersion, local overlap, decision-boundary geometry and architecture-specific feature use can also influence performance.

Image-based seed classification studies have reported useful predictive performance across several crop species [19–23]. However, predictive performance alone does not reveal the biological basis of discrimination. Here, DenseNet121 feature-space distances were examined alongside pairwise CNN F1 scores to provide complementary, image-derived descriptions of the five evaluated flax genotypes. These relationships are phenotypic rather than direct measures of genetic distance, and the ten pairwise comparisons share genotypes and are not statistically independent.

In particular, the fact that high-performing genotype pairs such as BeyazGelin–Culbert were also separated by greater distances in the feature space was consistent with an association between phenotypic differentiation and classifiability. In contrast, the closer positioning in the feature space of lower-performing comparisons such as Culbert–K0902 suggests that the model errors arise to a considerable extent from phenotypic similarity. This result reveals that low classification performance does not always indicate model inadequacy and may in some cases be a biologically expected outcome. CNN performance can therefore be evaluated not only as a technical success metric but also as an indirect indicator of the level of phenotypic separation among genotypes.

This interpretation is also supported by some studies that have emerged in recent years. Yang, Harrison, Zhang, Shen, Yan, Kang [24] reported that increasing genetic distance among different maize genotypes could affect the success of authenticity determination at the single-seed level. Similarly, Lim, Park, Baek, Botkin, Jang, Hong, Irish, Kim, Meinhardt, Curtin, Ahn [25] showed that the joint evaluation of the relationships among seed morphology, geographic origin and genetic structure contributed to a more accurate interpretation of the differences among cultivars. In the present study, the use of phenotypic distances calculated directly in the deep-learning-based feature space, rather than genetic distance, made it possible to reveal the biological patterns underlying classification success. In this respect, the study not only describes the differences among genotypes but also quantitatively evaluates the effect of these differences on classifiability.

Image-derived morphology and texture can be used to investigate phenotypic similarities across taxa, as illustrated by Lipiński [26]. Wang et al. [27] examined maize seed-coat morphology, color, and texture in phenotyping interaction networks. The similarity network in the present study likewise summarizes relationships among the evaluated flax genotypes in an image-derived feature space. It should not be interpreted as a genetic relatedness network.

The results obtained are also important for the explainability of deep-learning models. One of the main criticisms directed at artificial intelligence applications is that models are often regarded as “black boxes”. However, in this study, the demonstration of the relationship between phenotypic distance and classification performance supports the interpretation that model performance is related to measurable phenotypic separation, without establishing a causal or genetic mechanism. CNN models can therefore be evaluated not only as prediction tools providing high accuracy but also as phenotyping tools that help discover the relationships among genotypes. This approach considerably expands the role of deep learning in plant phenotyping studies.

CC: BY-NC-ND

Another notable point is that phenotypic distance is a continuous variable. Whereas in traditional classification approaches genotypes are mostly evaluated as correctly or incorrectly classified categories, the feature-space analyses allow the difference among genotypes to be examined gradationally. It is thus possible to evaluate not only whether two genotypes differ but also to what extent they differ. This approach can provide important advantages, particularly in applications such as parent selection, core-collection construction and genetic resource management.

Overall, the positive association between phenotypic distance and CNN classification performance represents an important finding of this study. The results indicate that image-based deep-learning models can encode phenotypic differences among the evaluated genotypes and that model performance may reflect, in part, their degree of phenotypic separation. Accordingly, classification performance can be considered alongside feature-space distance as complementary evidence when describing phenotypic similarity and separation among the evaluated genotypes. However, classification performance should not be interpreted as a direct or causal measure of phenotypic or genetic distance. Within the present experimental scope, this framework may support the characterisation of genetic resources, core-set evaluation and parent-selection research when used together with independent phenotypic and, ideally, molecular evidence.

4.4. Phenotypic core selection supports representative seed analysis

One of the most notable methodological aspects of this study is the construction, for each genotype, of phenotypic core seed sets. In conventional image-based classification studies, all of the samples in the dataset are generally used in model training. However, biological populations by their nature contain certain levels of variation, and important phenotypic differences can be observed even among individuals belonging to the same genotype. While some of these differences reflect natural variation, others may arise from environmental effects, developmental differences or measurement-induced deviations. As a result, some individuals in the datasets may have the character of outliers rather than representing the general phenotypic structure of the genotype. This situation can introduce noise, particularly into image-based classification models, making it more difficult to evaluate the true relationships among genotypes.

For this reason, rather than using all of the seeds, a core subset composed of the individuals closest to the phenotypic centre of each genotype was constructed in the study. The aim was to retain individuals that were more representative of the central phenotypic distribution while reducing the influence of atypical observations. This selection can make differences among the central phenotypic profiles of the evaluated genotypes easier to examine, but it does not establish that the retained 50% captures the complete or 'true' phenotypic structure of each genotype. In this respect, the phenotypic core approach can be regarded as an image-phenotyping adaptation of the representative-subset concept used in genetic resource studies rather than as an equivalent of a genetically defined core collection.

The core collection approach, long used in the field of genetic resource management, is based on a similar rationale. Selecting highly representative subsets from large collections both reduces evaluation costs and ensures that the fundamental variation structure of the collection is preserved. De Beukelaer et al. [3] reported that constructing core collections provides important advantages for the effective management of genetic diversity. Similarly, Franco, Crossa, Chen, Hearne [4] showed that sample-selection strategies have a direct effect on the representativeness of collections. The phenotypic core approach used in the present study can be regarded as an example of this concept adapted to the field of image-based phenotyping. Here, the aim is not to reduce the genetic material but to identify the phenotypically most representative individuals.

Another important result of phenotypic core selection is that within-genotype variation can be revealed more clearly. The dispersion analyses (Figure 7) showed that some genotypes exhibited more compact distributions than others, while some had a wider phenotypic spread. If all samples had been used without any filtering, individuals located far from the genotype centres could have affected the overall structure of the distribution and made it more difficult to interpret the distances among genotypes. Thanks to the core-seed approach, however, the characteristic phenotypic profile of each

CC: BY-NC-ND

genotype became more pronounced, and the relationships in the feature space could be evaluated more reliably.

This approach was also found to provide important advantages for the interpretability of deep-learning models. In recent years, achieving high accuracy alone has not been regarded as sufficient in phenotyping studies; it is also expected that the biological patterns on which the models base their decisions be understood. In datasets with high noise levels, models can sometimes be influenced by atypical individuals that do not represent the genotype, which can reduce the reliability of biological interpretations. In contrast, the phenotypic core approach, by ensuring that the analyses are carried out on individuals representing the genotype, allowed the results obtained from the feature-importance analyses, phenotypic distance calculations and similarity networks to be interpreted more consistently.

Phenotypic core selection also provided a common reference subset for visualising relationships among the evaluated genotypes. When the phenotypic distance analyses shown in Figure 6 are considered together with the similarity network presented in Figure 5, the feature space constructed from the core individuals shows structured differences among genotype pairs. Pairs located closer in this feature space tended, overall, to show lower classification performance, whereas more distant pairs tended to be more readily separated; however, pair-specific exceptions demonstrate that this relationship is not deterministic. These results suggest that the core datasets retain image-derived phenotypic patterns that are useful for comparative analysis, while not establishing that the learned feature space represents the complete biological or genetic structure of the genotypes.

In image-based seed phenotyping, the representativeness and quality of analyzed samples are important considerations alongside dataset size. Previous studies have examined morphological, colorimetric, and textural descriptors for soybean genotype classification [29], and the integration of image-derived and biochemical features for explainable classification of black cumin genotypes [30]. These studies support the utility of complementary phenotypic information but do not directly validate phenotypic-core selection. In the present study, selecting 200 seeds per genotype from 400 quality-controlled images provided a standardized subset for feature-based classification, CNN evaluation, and image-derived phenotypic comparisons. Because alternative core sizes and unfiltered datasets were not compared, the present results do not demonstrate improved classification accuracy or interpretability attributable to core selection. Its potential advantages require direct comparative validation.

Phenotypic core selection is therefore best interpreted as a representative-data strategy tested here at one predefined retention level (200 of 400 QC-passed seeds per genotype). The present experiment does not establish that 50% is an optimal core size, nor does it quantify how alternative retention levels would change classification performance or inter-genotype distances. Future work should explicitly compare several core proportions (for example 25%, 50% and 75%) and evaluate stability of both predictive metrics and feature-space relationships before general recommendations are made.

4.5. Implications for germplasm characterization and breeding

The findings of this study indicate that image-based deep-learning approaches may provide useful information not only for genotype discrimination but also for the phenotypic characterization of genetic resources. Effective identification of plant genetic resources and accurate assessment of phenotypic diversity are important components of both germplasm conservation and breeding programs [1, 3]. In this context, quantitative information extracted from single-seed images provides a rapid, reproducible, and non-destructive means of characterizing phenotypic differences among genotypes. The combination of pairwise classification, deep feature-space distances and genotype relationship networks further shows that seed images can provide information beyond categorical identification by describing the relative phenotypic relationships among the materials examined.

Such phenotypic information may also complement decision-making in plant breeding. Parent selection is a critical step affecting the generation of useful variation and subsequent genetic gain [4].

CC: BY-NC-ND

Although pedigree information, agronomic performance and molecular data remain central to this process, image-derived phenotypic distances could provide an additional layer of information for rapidly screening differences among candidate parental materials. Genotypes that are more widely separated in the learned feature space can be regarded as more divergent in terms of the seed phenotypes captured by the imaging system. However, this divergence should not be interpreted as evidence of greater genetic distance or superior crossing potential. Rather, image-derived phenotypic distance may serve as a complementary screening criterion to be considered together with agronomic, pedigree and molecular information when available.

Image-based genotype discrimination may also have potential applications in varietal-purity assessment and seed quality-control systems. Maintaining varietal purity is important for the consistency and quality of agricultural production, and automated imaging systems offer the possibility of screening large numbers of seeds rapidly and non-destructively [7]. The pairwise classification results obtained here demonstrate that genotype-specific seed phenotypes can be discriminated with varying degrees of success, including some phenotypically similar genotype pairs. These findings provide a basis for further investigation of image-based approaches for detecting varietal mixtures and supporting seed-certification workflows. Such applications, however, would require dedicated validation using intentionally mixed seed lots, independent production environments and substantially larger genotype panels before operational use could be considered.

The phenotypic-core approach additionally highlights the importance of sample composition in digital phenotyping. Selecting seeds located closest to their genotype-specific centroid provided a standardized subset representing the central region of each genotype's observed phenotypic distribution. This strategy enabled subsequent classification and feature-space analyses to be conducted on phenotypically representative subsets, but it should not be interpreted as demonstrating that a 50% retention level is optimal or that exclusion of more peripheral phenotypes necessarily improves model performance. Future studies should therefore compare alternative core sizes and selection strategies and, where appropriate, evaluate models trained on core subsets against models using the complete phenotypic distribution.

Several limitations define the scope of these findings. The study comprised five flax genotypes—three advanced breeding lines (EBYC01, K0902, and K1104) and two control cultivars (BeyazGelin and Culbert)—evaluated over two growing seasons. Consequently, the observed classification patterns and phenotypic relationships are specific to these materials and should not be generalized to the broader phenotypic diversity of flax germplasm without validation in larger and more diverse panels. Scanner-based imaging under standardized laboratory conditions minimized acquisition variability but does not establish transferability to field imaging, mobile-camera systems or other acquisition platforms. Moreover, DenseNet121 feature-space distance represents image-derived phenotypic divergence and must not be interpreted as genetic distance. The phenotypic core was constructed using a single 50% retention level before the train/validation/test split; therefore, the reported model performance pertains specifically to the selected core population, while the effects of alternative retention levels remain unresolved. The CNN analyses also used transfer learning on relatively small pairwise image sets (200 core images per genotype before splitting). Pretrained weights, data augmentation, validation monitoring and restricted fine-tuning reduce but do not eliminate the risk of overfitting, and the present study does not establish external generalization. Independent external datasets, additional genotypes and environments, and repeated acquisition with different imaging systems are therefore required before deployment-oriented claims can be made. Finally, the significant association between DenseNet121 feature-space distance and CNN classification performance was derived from only ten genotype pairs and should be interpreted as an association rather than evidence of causation.

Taken together, the results support an integrated framework combining representative seed selection, quantitative seed phenotyping, machine-learning and deep-learning classification, and deep feature-space analysis. Within the biological and experimental scope examined here, this framework provides complementary information on genotype discrimination and phenotypic relationships. Validation across larger germplasm collections, multiple environments and independent imaging systems will be necessary to determine whether the approach can be generalized to broader flax

CC: BY-NC-ND

germplasm and, subsequently, to other crop species.

5. Conclusion

This study demonstrates that standardized seed imaging, representative phenotypic-core selection, feature-based machine learning and CNN feature-space analysis can be combined to characterize pairwise phenotypic relationships among the five evaluated flax genotypes. Classification performance varied among genotype pairs, and DenseNet121 centroid distance was positively associated with CNN F1 score, although pair-specific exceptions showed that distance alone does not determine discriminability.

The phenotypic-core strategy provided a common representative subset for comparing handcrafted descriptors, CNN classifications, and feature-space analyses. Within this design, genotype-specific dispersion and inter-genotype centroid distance supplied complementary descriptions of the selected seed populations. These quantities are phenotypic descriptors of standardized seed images and should not be interpreted as direct measures of genetic distance or as evidence that the selected 50% core is universally optimal.

The framework may be useful as a non-destructive screening and characterization approach for seed phenotyping, but broader generalization requires validation with more genotypes, independent environments, alternative imaging systems, and explicit core-size sensitivity analyses. Future studies integrating molecular information with image-derived phenotypic representations would be particularly valuable for determining how far learned feature-space relationships correspond to underlying genetic relationships.

Peer-review: External, Independent.

Acknowledgements:
The author would like to thank the Trakya Agricultural Research Institute for providing the flax germplasm used in this study and the Farmer Training Branch Directorate (BAÇEM) of Balıkesir Metropolitan Municipality for facilitating the field experiments.

Declarations:

1. Statement of Originality:
The author declares that this manuscript presents original research and has not been published previously or submitted simultaneously for publication elsewhere.

2. Author Contributions:
Concept: SA; ***Conceptualization:*** SA; ***Literature Search:*** SA; ***Data Collection:*** SA; ***Data Processing:*** SA; ***Analysis:*** SA; ***Writing – original draft:*** SA; ***Writing – review & editing:*** SA.

3. Ethics approval:
Not applicable.

4. Funding/Support:
This work has not received any funding or support.

5. Competing Interests:
The author declares no competing interests.

6. GenAI Usage Statement:
Generative artificial intelligence tools were used to assist with language editing and manuscript revision

7. Sustainable Development Goals:

CC: BY-NC-ND

REFERENCES

[1] Diederichsen, A., & Raney, J. P. (2006). Seed colour, seed weight and seed oil content in Linum usitatissimum accessions held by Plant Gene Resources of Canada. Plant Breeding, 125(4), 372–377. https://doi.org/10.1111/j.1439-0523.2006.01231.x

[2] Walkowiak, M., Spasibionek, S., & Krótka, K. (2022). Variation and genetic analysis of fatty acid composition in flax (Linum usitatissimum L.). Euphytica, 218, Article 2. https://doi.org/10.1007/s10681-021-02941-6

[3] De Beukelaer, H., Davenport, G. F., & Fack, V. (2018). Core Hunter 3: Flexible core subset selection. BMC Bioinformatics, 19, 203. https://doi.org/10.1186/s12859-018-2209-z

[4] Franco-Duran, J., Crossa, J., Chen, J., & Hearne, S. J. (2019). The impact of sample selection strategies on genetic diversity and representativeness in germplasm bank collections. BMC Plant Biology, 19, 520. https://doi.org/10.1186/s12870-019-2142-y

[5] Deng, Q., Zhao, S., Guan, C., Fu, G., Yu, S., Lin, S., Wang, Z., Fu, H., Lu, X., & Cheng, S. (2024). Construction and evaluation of pepper core collection based on phenotypic traits and SSR markers. Applied Sciences, 14(17), 7473. https://doi.org/10.3390/app14177473

[6] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

[7] Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology, 37, 38–44. https://doi.org/10.1038/nbt.4314

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

[9] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[12] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243

[13] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and

CC: BY-NC-ND

Pattern Recognition (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

[14] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (pp. 6105–6114). PMLR.

[15] Wiesnerová, D., & Wiesner, I. (2008). Computer image analysis of seed shape and seed color for flax cultivar description. Computers and Electronics in Agriculture, 61(2), 126–135. https://doi.org/10.1016/j.compag.2007.10.001

[16] Smykalova, I., Grillo, O., Bjelkova, M., Pavelek, M., & Venora, G. (2013). Phenotypic evaluation of flax seeds by image analysis. Industrial Crops and Products, 47, 232–238. https://doi.org/10.1016/j.indcrop.2013.03.001

[17] Kang, Q. H., Sun, Z. Y., Qiu, C. S., Wang, Y. F., Zhu, X., Song, X. X., Chen, X. Y., Yao, D. D., Jiang, W. D., Yuan, H. M., Tang, L. L., Liu, D. D., & Yao, Y. B. (2024). The study progress on breeding of flax and industrial hemp in China. Journal of Natural Fibers, 21(1), 2389159. https://doi.org/10.1080/15440478.2024.2389159

[18] Lu, J., Wu, H., Wang, F., Li, J., Wang, Y., Zhao, Q., Wang, Y., Wang, X., Lei, X., Sun, R., Zhang, J., Xiong, A., Deyholos, M. K., & Zhang, J. (2025). Telomere-to-telomere flax (Linum usitatissimum L.) genome assembly unlocks insights beyond fatty acid metabolism pathways. Horticulture Research, 12, uhaf127. https://doi.org/10.1093/hr/uhaf127

[19] Xu, P., Yang, R., Zeng, T., Zhang, J., Zhang, Y., & Tan, Q. (2021). Varietal classification of maize seeds using computer vision and machine learning techniques. Journal of Food Process Engineering, 44(11), e13846. https://doi.org/10.1111/jfpe.13846

[20] 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. https://doi.org/10.3390/plants10071406

[21] Yang, H., Ni, J., Gao, J., Han, Z., & Luan, T. (2021). A novel method for peanut variety identification and classification by improved VGG16. Scientific Reports, 11, 15756. https://doi.org/10.1038/s41598-021-95240-y

[22] Akbarpour, O., Akbari, S., Fanourakis, D., & Taheri-Garavand, A. (2026). Deep learning-based seed variety classification: A case study in maize. BMC Plant Biology, 26, 105. https://doi.org/10.1186/s12870-025-07919-3

[23] Wei, J., Dai, Z., Zhang, Q., Yang, L., Zeng, Z., Zhou, Y., Liu, J., & Chen, B. (2025). Seed multispectral imaging combined with machine learning algorithms for distinguishing different varieties of lettuce (Lactuca sativa L.). Food Chemistry: X, 27, 102399. https://doi.org/10.1016/j.fochx.2025.102399

[24] Yang, Y., Harrison, R. C., Zhang, D., Shen, B., Yan, Y., & Kang, D. (2024). Effect of genetic distances of different genotypes of maize on the authenticity of single seeds detected by NIR spectroscopy. Frontiers in Plant Science, 15, 1361328. https://doi.org/10.3389/fpls.2024.1361328

[25] Lim, S., Park, S., Baek, I., Botkin, J. R., Jang, J. H., Hong, S. M., Irish, B., Kim, M. S., Meinhardt, L., Curtin, S. J., & Ahn, E. (2025). Integrative analysis of seed morphology, geographic origin, and genetic structure in Medicago with implications for breeding and conservation. BMC Plant Biology, 25, 274. https://doi.org/10.1186/s12870-025-06304-4

[26] Lipiński, S. (2025). Statistical analysis of seeds morphology and texture for interspecific similarity assessment across taxonomic levels. Biuletyn Instytutu Hodowli i Aklimatyzacji Roślin, 303/304, 19–33. https://doi.org/10.37317/biul-2025-0006

CC: BY-NC-ND

[27] Wang, Y., Zhang, Y., Song, A., Guo, M., Zhang, C., Wang, C., Zhao, Y., Huang, G., Men, Q., Zhao, C., & Guo, X. (2025). Precision profiling of seed coat phenotypes in maize: 3D surface morphology, color, texture traits for the construction of phenotyping interaction networks. Current Plant Biology, 43, 100493. https://doi.org/10.1016/j.cpb.2025.100493

[28] Taghavi Namin, S., Esmaeilzadeh, M., Najafi, M., Brown, T. B., & Borevitz, J. O. (2018). Deep phenotyping: Deep learning for temporal phenotype/genotype classification. Plant Methods, 14, 66. https://doi.org/10.1186/s13007-018-0333-4

[29] Akçura, S. (2026). Image-based multimodal seed phenotyping using morphological, colorimetric, and textural traits for soybean genotype classification. Genetic Resources and Crop Evolution, 73(6), 232. https://doi.org/10.1007/s10722-026-02864-y

[30] Akçura, S. (2026). Explainable multimodal learning for genotype classification using image-based and biochemical features: A case study on black cumin (Nigella sativa L.). Food Analytical Methods, 19(10), 291. https://doi.org/10.1007/s12161-026-03202-9

CC: BY-NC-ND