AGROLOGY

ISSN 2617-6106 (print)
ISSN 2617-6114 (online)
Agrology, 2024, 7(3), 112–118
doi: 10.32819/202415

Original researches

Received: 27.06.2024
Revised: 01.08.2024
Accepted: 16.08.2024

Dnipro State Agrarian
and Economic University,
Serhiya Yefremova st., 25,
Dnipro, 49600, Ukraine.
Tel.: +38-068-861-44-37.
E-mail: aliev@meta.ua

Institute of Oilseed Crops
of the National Academy
of Agrarian Sciences of Ukraine,
Instytutska st., 1,
village Sonyachne, 69055,
Zaporizhia region, Ukraine.
Tel.: +38-050-258-16-17.
E-mail:
ved-medeva.katerina@gmail.com

Cite this article: Aliiev, E.,
& Vedmedeva, K. (2024).
Systematization of sunflower
genotypes based on seed pheno-
typic characteristics using neural
networks. Agrology, 7(3), 112–
118. doi: 10.32819/202415

Systematization of sunflower genotypes based on seed phenotypic characteristics using neural networks

E. Aliiev*, K. Vedmedeva**
**Dnipro State Agrarian and Economic University, Dnipro, Ukraine*
***Institute of Oilseed Crops of the National Academy of Agrarian Sciences of Ukraine, Sonyachne, Ukraine*

Abstract. In recent years, various datasets related to the phenotyping of sunflower genotypes have become increasingly accessible. However, one of the key challenges remains the efficient and accurate prediction of phenotypes based on genotypes in the context of climate change. Analyzing phenotypes at different levels of organization and detecting connections between phenotypes and genotypes require the integration and processing of large, diverse, and often noisy datasets. Machine learning offers a broad arsenal of methods and approaches for identifying predictive patterns in such data. Therefore, the research aimed to develop a methodology for the systematization of sunflower genotypes based on seed phenotypic characteristics using the data vector quantization method and neural networks. The study revealed the phenotypic characteristics of sunflower seeds from various genotypes selected by the Institute of Oilseed Crops of NAAS, grown in the southern Steppe of Ukraine, including seed length, width, thickness, seed mass, kernel mass, and seed coat cracking force. For this purpose, appropriate laboratory equipment was developed, including two modules for determining the morphological and rheological properties of seeds. The developed methodology for the systematization of sunflower genotypes based on seed phenotypic characteristics includes the following steps: measuring the characteristics of sunflower seeds from various samples (parental components); studying the mutual correlation of characteristics; conducting hierarchical cluster analysis of the data using the Ward's method; determining the optimal number of groups; performing k-means clustering using the vector quantization method; determining the correspondence of ranges of characteristics to the group; training a neural network to group the data by samples and created groups; verifying the adequacy of the neural network on test data. The developed methodology was tested, and the MLP 30-15-3 neural network for grouping data by samples and created groups of sunflower seeds was developed in the Statistica software package. The network's training efficiency was 99.4%, and such of testing and validation was 95.6% and 96.7%, respectively.

Keywords: sunflower; seeds; phenotype; characteristics; correlation; grouping; neural network.

Introduction

In recent years, the availability of various datasets related to plant genotype phenotyping has increased (Bayer et al., 2021; Najafabadi et al., 2021; Van-Dijk et al., 2021). However, a primary challenge in modern plant science is the efficient and accurate prediction of phenotypes based on genotypes in the context of climate change. Genetic variations influence biochemical processes in cells and tissues, which, along with environmental signals, determine organ development, plant growth, yield, and resistance to abiotic and biotic factors. In the context of modern plant breeding, studying the impact of environmental and genotypic changes opens new horizons in regulating key processes in the plant life cycle. At the same time, this also poses challenges regarding the accuracy of predicting crop yields under the changing climate conditions.

Analyzing phenotypes at various levels of organization and developing connections between phenotypes and genotypes require the integration and processing of large, diverse, and noisy datasets. Machine learning offers a wide range of methods and approaches for identifying predictive patterns in such data. Classification programs based on machine learning are becoming powerful tools for creating accurate and reliable models used to assess the quality of agricultural products. These programs include various algorithms, such as decision trees, artificial neural networks, genetic algorithms, regression, and fuzzy logic. These algorithms are useful for developing machine learning models that assist in selecting important features and correcting complex input-output mapping approaches (Libbrecht & Noble, 2015; Crossa et al., 2017; Perez-Sanz et al., 2017).

Sunflower (*Helianthus annuus* L.) is known for its heterosis expression across multiple traits, making it an interesting research subject in terms of adaptive introgression and evolutionary biology. Sunflower is valued as a model plant for studying flower development processes and as a source of premium-quality edible oil. Currently, sunflower is the fourth most important oilseed crop globally, thanks to its ability to grow in various conditions and high yield. Often, hybrids (F1) derived from distantly related high-yielding lines outperform their parent lines in productivity (Ibrar et al., 2022; Ibrar et al., 2024).

Machine learning already plays a significant role in various engineering fields, but its potential in applied plant breeding, particularly for sunflowers, is still underexplored (Aliiev, 2023). Few studies focus on the systematization of plant genotypes across different crops based on phenotypic characteristics using neural networks.

A detailed review of the assessment and classification of plant traits during phenotyping using machine vision methods was conducted by Kolhar & Jagtap (2023). This paper provides a comprehensive review of modern machine vision approaches for analyzing and classifying plant traits. The available datasets enabling unified comparison of current phenotyping methods were discussed, and promising research directions involving deep learning-based machine vision algorithms for structural (2-D and 3-D), physiological, and temporal trait evaluation and classification studies in plants were highlighted.

In the study by Jin et al. (2022), a multi-objective approach for classifying sunflower seeds based on sparse convolutional neural networks was developed. Sunflower seeds were obtained from a photo image created using the YOLOv5 object detection algorithm, and a ResNet-based model was used for their classification by external traits.

The drawbacks of ResNet include a large number of parameters and high memory requirements. Rajalakshmi et al. (2022) developed the RiceSeedNet deep neural network for identifying rice seed varieties. The Vision Transformer-based architecture named RiceSeedNet was developed to automate the process of identifying rice varieties. The RiceSeedNet model was also tested on a publicly available rice grain dataset to evaluate its performance in classifying different rice varieties.

In the Eldem's (2020) study, a new deep neural network (DNN) model for wheat seed classification was proposed, with data taken from the UCI Machine Learning Repository. The model was tested on various combinations of training and test data using both UCI data and synthetically generated sets. As a result, 100% classification accuracy was achieved. The proposed model outperformed other studies presented in the literature for wheat classification.

Luan et al. (2020) developed a CNN model with eight convolutional layers for image feature extraction, and a skip connection was used to enhance the model's learning capability. Compared to the classical convolutional network, it has a smaller size without reducing accuracy. However, due to the CNN's characteristics, the high-dimensional features obtained by the convolutional layers are rarely explored due to their high dimensionality and lack of global representation. Large-scale experiments demonstrate that the developed model with an attention mechanism achieves better accuracy on the sunflower seed image dataset compared with several classical networks.

Li et al. (2020) carried out a detailed review of the main plant phenotyping methods based on computer vision, including their principles, application areas, results, and comparisons. The authors analyzed over 200 studies of plant phenotyping, focusing on their technical evolution over more than twenty years (from 2000 to 2020). The review covers a wide range of topics, such as imaging technologies, plant datasets, and modern phenotyping methods. Plant phenotyping is divided into two main categories: individual plant organ phenotyping and whole plant phenotyping.

The paper by Gao et al. (2024) presented the Crop-GPA, a comprehensive and functional platform for analyzing gene-to-phenotype associations in agricultural crops. The current version of Crop-GPA provides researchers with carefully curated information about genes, phenotypes, and their associations (GPA) through a user-friendly interface, interactive graphical visualizations, and powerful online tools. Two computational tools, GPA-BERT and GPA-GCN, were specifically developed and integrated into Crop-GPA to automatically extract gene-to-phenotype associations from agricultural literature and predict unknown links based on known associations.

The presented artificial neural networks are sufficiently individual for each crop and have closed access, significantly complicating their use for personal research.

Therefore, the research aims to develop a methodology for the systematization of sunflower genotypes based on seed phenotypic characteristics using the data vector quantization method and neural networks.

Materials and methods

The object of the research is the patterns of phenotypic characteristics of sunflower genotypes. The selected phenotypic characteristics of sunflower seeds in the seed head are as follows: geometric dimensions (length L, width W, thickness T), seed mass (Ms), kernel mass (Mk), and seed coat cracking force (F).

The sunflower samples were chosen from the collection of the Institute of Oilseed Crops of NAAS, selected based on morphological and marker traits. The samples included the following: KP11A, KP11, 340V, 162V, 165ar2, 164h, 168v, ZL169A, ZL54A, 178a, 168b, 174d, 175bp1, 174e, 168b, 168a, 178a×238, ZL50A, SKH75A, 160v, 162bp1, Bilochka, Prometey, and Zaporiz′kyy Kondyters′kyy.

The samples were grown on the experimental plots of the Institute of Oilseed Crops of NAAS in two different conditions (main field and nursery) in 2023. The soil of the experimental plots was regular, medium-power, low-humus black soil. The natural and climatic conditions corresponded to the southern Steppe of Ukraine. Overall, the weather conditions during the 2023 growing season were satisfactory for sunflower growth and development. The main difference in the nursery was the presence of a concrete fence, which reduced ventilation of the crops and created additional adverse microclimatic conditions, such as the development of fungal diseases at all stages and increased the surface soil temperature.

Table 1
Soil and climatic conditions of the experimental sites (x ± SE)

Indicator	Main field	Nursery
Soil composition		
Humus content in the arable layer up to 30–40 cm, %	3.3 ± 0.2	3.0 ± 0.2
Available nitrogen content, mg/100 g of soil	7.8 ± 0.6	7.0 ± 0.4
Mobile phosphorus content, mg/100 g of soil	9.9 ± 0.4	9.3 ± 0.3
Mobile potassium content, mg/100 g of soil	16.0 ± 0.7	15.5 ± 0.4
Soil pH	6.7 ± 0.3	6.7 ± 0.3
Natural and climatic conditions		
Average daily temperature during the growing season, °C	22.3 ± 0.7	22.3 ± 0.7
Total precipitation during the growing season, mm	172.1 ± 2.3	165.0±2.1*

Note: * – P < 0.05.

The plot size for each sample was 14 m², in three repetitions. Each sample included 15 plants. Seed sampling from the sunflower seed heads was conducted in a spiral pattern, with five spirals subjected to analysis (repeated measurements).

The research of the geometric dimensions of the sunflower seeds was carried out using specially designed laboratory equipment – a module for determining the morphological characteristics of seeds (Fig. 1). The module's algorithm was based on the image analysis method (Aliiev, 2020). This method reduces the time required for seed preparation and image acquisition. The contours of the seeds were automatically detected on digital images, and several shape parameters were calculated, including length (L), width (W), and thickness (T).

The seed coat cracking force was identified using specially designed laboratory equipment – a module for determining the rheological properties of seeds (Fig. 2). The measurement algorithm for determining the force involved several stages:

1. Stage One. The seed was placed in the required position on a flat horizontal surface of the working table, directly under a cylindrical indenter (20 mm in diameter), which was attached to a load cell moving in the vertical plane. The load cell with the attached indenter must be calibrated, and in the resting state, it should register a force of 0 N.

2. Stage Two. The indenter begins to descend uniformly and linearly from the top position at a speed of 10 mm/s. As soon as the indenter contacts the seed, the force on the load cell begins to increase from 0 N. This moment is recorded as the start of the measurement process.

3. Stage Three. During the measurement process, the compressive force Ft is determined. When the value of Ft begins to decrease, the value recorded at that moment is the seed coat cracking force F.

The seed mass M is measured using a load cell with strain gauges, the HX711 module, and an Arduino control board, which are part of the morphological properties measurement module. The strain gauges and HX711 module are connected in a bridge configuration. The HX711, produced by Avia Semiconductors, is a 24-bit analog-to-digital converter (ADC) with an integrated operational amplifier.

Statistical processing of the obtained data and the development of the neural network were conducted using the Statistica software package (StatSoft).

Results

As a result of measuring seeds from different samples in the main field, a dataset of average values and deviations of their primary characteristics was obtained. The analysis of these data was carried out in several stages.

The first stage involved studying the mutual correlation between sunflower seed characteristics from various samples (Table 2). To assess the validity of the relationship between two characteristics, the Student's t-test was calculated (Table 2) and compared with its tabulated value trt (0.05; 187) = 1.973. Analyzing the obtained data, it can be stated that there was a moderate correlation between the characteristics (according to the Cheddock scale). The correlation between seed coat cracking force and other characteristics was not significant (ranging from 0.0005 to 0.0179). However, there was a certain correlation between geometric dimensions and seed and kernel mass, ranging 0.6355 to 0.7528 and 0.5580 to 0.6918, respectively. This correlation was quite logical and consistent with other studies (Nosal et al., 2017; Vedmedeva & Nosal, 2020; Jin et al., 2022).

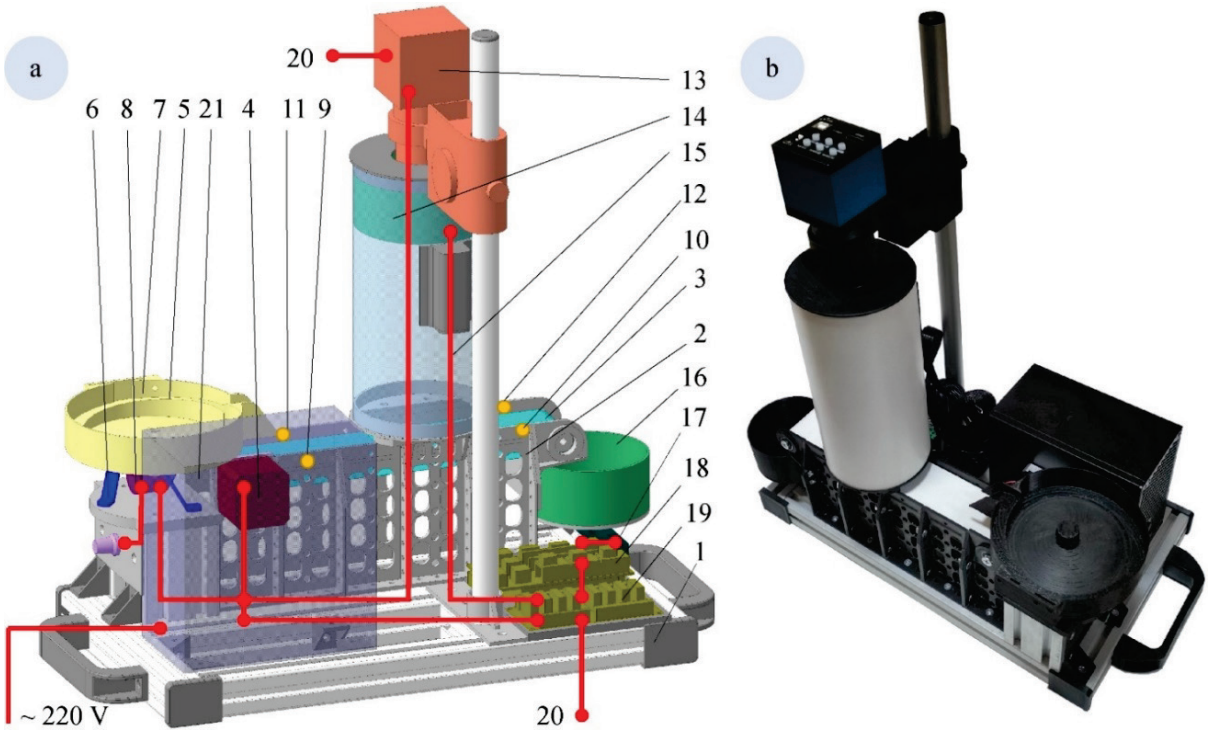

Fig. 1. Constructive-technological scheme (*a*) and general view (*b*) of the morphological properties measurement module: *1* – frame; *2* – belt conveyor; *3* – belt; *4* – electric motor; *5* – seed feed tray; *6* – dampers; *7* – shutter; *8* – vibration motor; *9* – infrared leds; *10* – infrared leds; *11* – photo receivers; *12* – photo receivers; *13* – camera; *14* – RGB-leds; *15* – lightproof tube; *16* – receiving tray; *17* – load cell; *18* – amplifier; *19* – control unit; *20* – personal computer; *21* – power supply

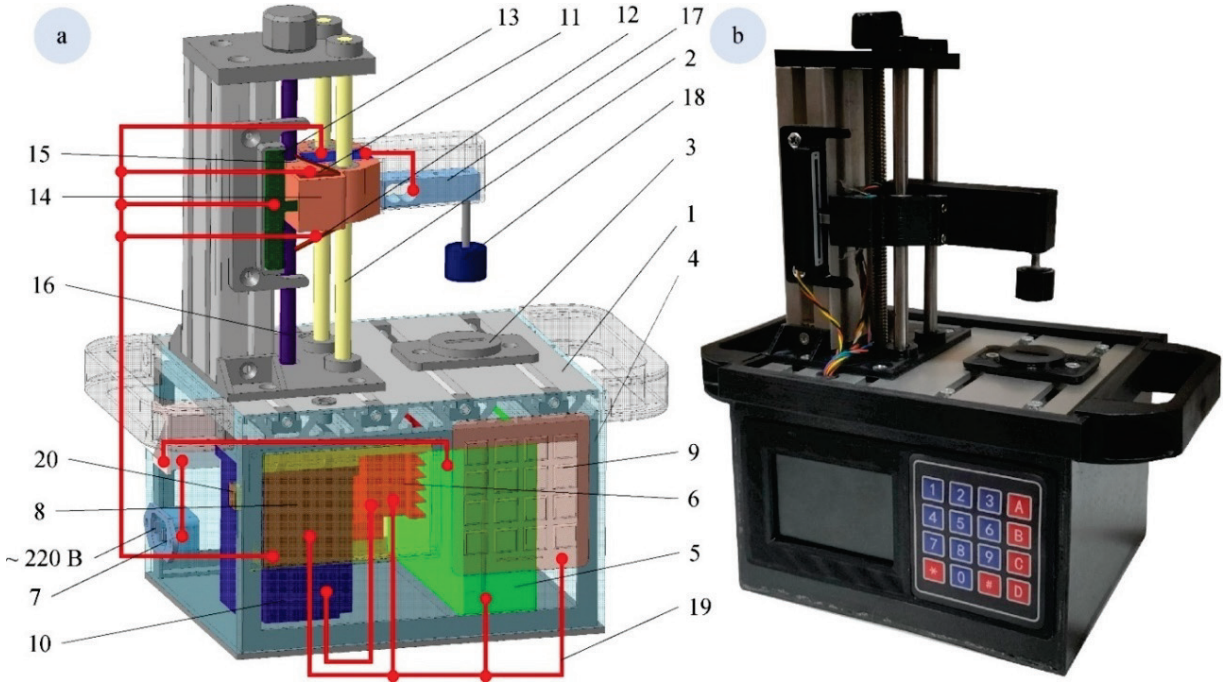

Fig. 2. Constructive-technological scheme (*a*) and general view (*b*) of the rheological properties measurement module: *1* – base; *2* – guide rail; *3* – work table; *4* – body; *5* – power supply unit; *6* – stepper motor controller (TB6600); *7* – socket with switch; *8* – control unit with LCD display (Arduino MEGA 2560 + 3.5 Inch TFT Color Display 320x480 Screen); *9* – keyboard; *10* – stepper motor; *11* – upper limit switch; *12* – lower limit switch; *13* – linear variable resistor; *14* – correction; *15* – nut; *16* – screw; *17* – load cell; *18* – indenter; *19* – electrical wires; *20* – USB output for connection to personal computer; *21* – amplifier (Weight Sensor HX711)

Based on the Ward's method, the second stage involved performing hierarchical cluster analysis of the data to determine the number of groups into which the data can be divided. The Ward's minimum variance criterion minimizes the total variance within clusters. To implement this method, at each step, the pair of clusters is found that results in the smallest increase in the total within-cluster variance after merging. This increase is a weighted squared distance between the cluster centers. Initially, all clusters are singletons (clusters containing one point). To apply the recursive algorithm for this objective function, the initial distance between individual objects must be proportional to the square of the Euclidean distance. Thus, the initial distances in the Ward's minimum variance method are determined as the square of the Euclidean distance between points (de Amorim, 2015; Weylandt et al., 2019; Brusco et al., 2024). To visualize the results of hierarchical clustering, we constructed a dendrogram (Fig. 3).

Table 2
Graphical visualization and calculated correlation coefficients (r) and Student's t-test (t_r) of mutual correlation of sunflower seed characteristics from different samples

Characteristic	Length L, mm	Thickness T, mm	Width W, mm	Seed mass M_s, g	Kernel mass M_k, g	Cracking force F, N
Length L, mm		r = + 0.7528 $t_r = 9.124 > t_r^t$	r = + 0.6355 $t_r = 5.921 > t_r^t$	r = + 0.7070 $t_r = 8.035 > t_r^t$	r = + 0.4593 $t_r = 7.646 > t_r^t$	r = – 0.0179 $t_r = 0.755 < t_r^t$
Thickness T, mm			r = + 0.7147 $t_r = 5.002 > t_r^t$	r = + 0.7768 $t_r = 8.079 > t_r^t$	r = + 0.5060 $t_r = 7.393 > t_r^t$	r = – 0.0045 $t_r = 0.114 < t_r^t$
Width W, mm				r = + 0.6918 $t_r = 5.771 > t_r^t$	r = + 0.4922 $t_r = 9.132 > t_r^t$	r = – 0.0005 $t_r = 0.587 < t_r^t$
Seed mass M_s, g					r = + 0.5580 $t_r = 10.824 > t_r^t$	r = + 0.0102 $t_r = 0.781 < t_r^t$
Kernel mass M_k, g						r = + 0.0336 $t_r = 0.210 < t_r^t$
Cracking force F, N						

Note: on the graphs, the x-axis corresponds to the column names of the Table, while the y-axis corresponds to the row names of the Table; blue dots represent experimental data, and the red line indicates the linear regression line between the two seed characteristics; histograms show the distribution of seed characteristics corresponding to the column and row of the table; the statistical significance level is P < 0.05; trt (0.05; 187) = 1.973 is the tabulated value of the Student's t-test.

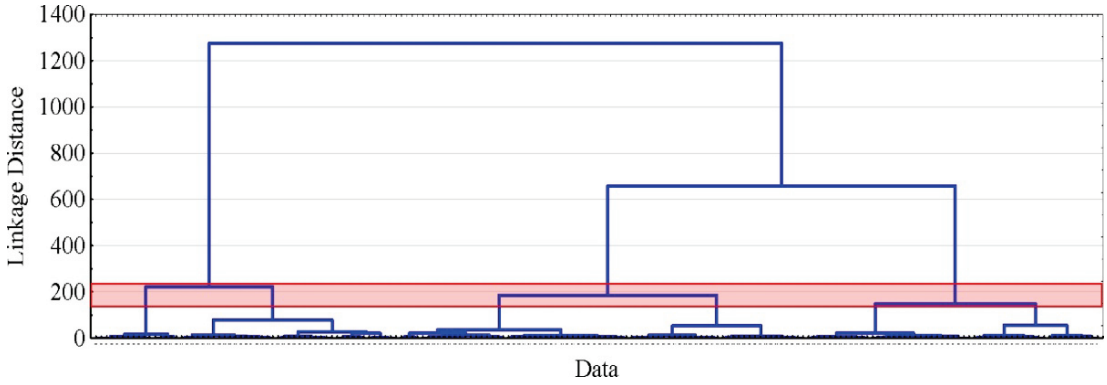

Fig. 3. Dendrogram of distances between steps in hierarchical clustering results

It shows the degree of proximity between individual objects and clusters and visually demonstrates the sequence of their merging or splitting. The number of levels in the dendrogram corresponds to the number of steps in the merging or splitting of clusters. The number of branches corresponds to the possible number of groups. We see that at a linkage distance of 170–240, the branches are at the same level, so we accept the number of groups as 3.

Next, in the third stage, k-means clustering was performed, which is a vector quantization method aimed at partitioning n observations into k groups, where each observation belongs to the cluster with the nearest mean–cluster centers or centroids (Hamerly & Elkan, 2004; Amorim & Hennig, 2015; Pasi, 2018). The mean values for each seed characteristic and statistical indicators are presented in Table 3. The tabulated Fisher criterion ft (6, 188) = 6.88 is less than the calculated value, which confirms the adequacy of the obtained data processing results. After ranking the means for each group within the range from 0 to 1 (where 0 is the smallest value and 1 is the largest value), a graphical visualization of the obtained grouping was created (Fig. 4).

Multifactorial analysis of variance of single-factor data revealed differences between groups and had sufficient power (Table 4). Thus, it can be stated that the classification of the data into three groups is adequate. The results of data clustering are presented in Figure 5.

The three-dimensional dependencies clearly demonstrate the separation of samples into three groups. Group 1 includes the following samples: KP11A, KP11, 340V, 162V, 165ar2, 164g, 168v. Group 2 includes: ZL169A, ZL54A, 178a, 168b, 174d, 175bp1, 174e, 168b, 168a, 178a × 238. Group 3 includes: ZL50A, SKH75A, 160v, 162bp1, Bilochka, Prometey, Zaporiz′kyy kondyters′kyy, and 165Vr1. This grouping of the indicated samples was preliminarily confirmed by the results of the studies (Nosal et al., 2017; Vedmedeva & Nosal, 2020).

The fourth stage was training the neural network (Neural Networks) for data clustering based on samples and created groups. The neural network consists of interconnected blocks or nodes called artificial neurons, which loosely model brain neurons. They are connected by edges that model synapses in the brain. Each artificial neuron receives signals from connected neurons, processes them, and sends a signal to other connected neurons. The signal is a real number, and the output of each neuron is calculated using a nonlinear function of the sum of its inputs, known as the activation function. The strength of the signal at each connection is determined by a weight that is adjusted during training (Bishop, 2006).

Table 3
Mean values (x ± SE) of groups for each sunflower seed characteristic and calculated statistical parameters

Characteristics / Indicator	Length L, mm	Thickness T, mm	Width W, mm	Seed mass M_s, g	Kernel mass M_k, g	Breaking force F, N
Group 1	8.3 ± 0.8	3.5 ± 0.4	3.9 ± 0.4	0.092 ± 0.021	0.069 ± 0.018	49.1 ± 8.1
Group 2	6.6 ± 0.9	2.8 ± 0.4	3.2 ± 0.3	0.052 ± 0.019	0.039 ± 0.015	31.1 ± 9.6
Group 3	9.6 ± 0.7	4.2 ± 0.4	4.8 ± 0.5	0.019 ± 0.016	0.019 ± 0.009	13.3 ± 8.9
Intergroup variance	66.6	32.4	28.5	0.070	0.012	34832
Intragroup variance	259.4	93.5	137.8	0.274	0.161	5669
Fisher's test F	23.7	32.1	19.1	22.1	12.03	568.3

Note: the level of statistical significance P < 0.05.

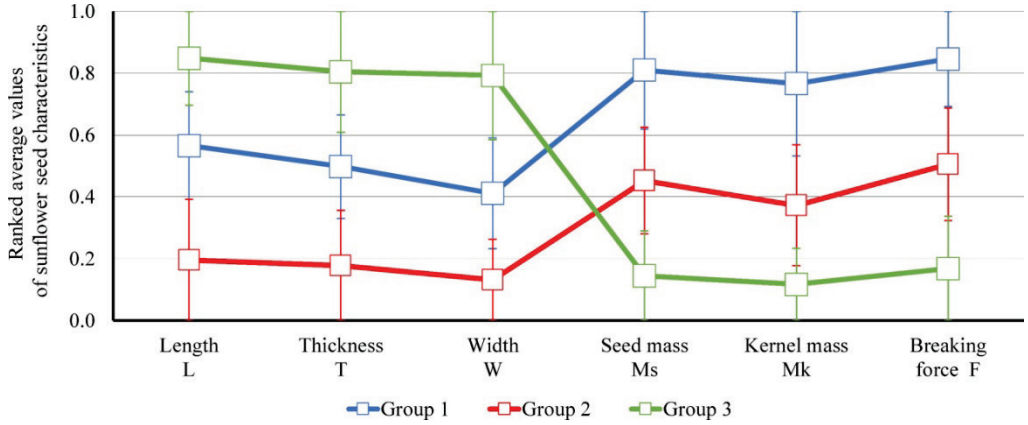

Fig. 4. Graph of ranked mean values of groups for each sunflower seed characteristic

Table 4
Multifactorial significance tests, effect sizes, and power for the Wilks test

Test	Value	F	Effect	Error	P	Partial eta-squared	Non-centrality	Observed power (P = 0.05)
Intercept	0.006	4879.3	6	180	<0.001	0.993	29276.2	1.000
Groups	0.109	60.4	12	360	<0.001	0.668	725.5	1.000

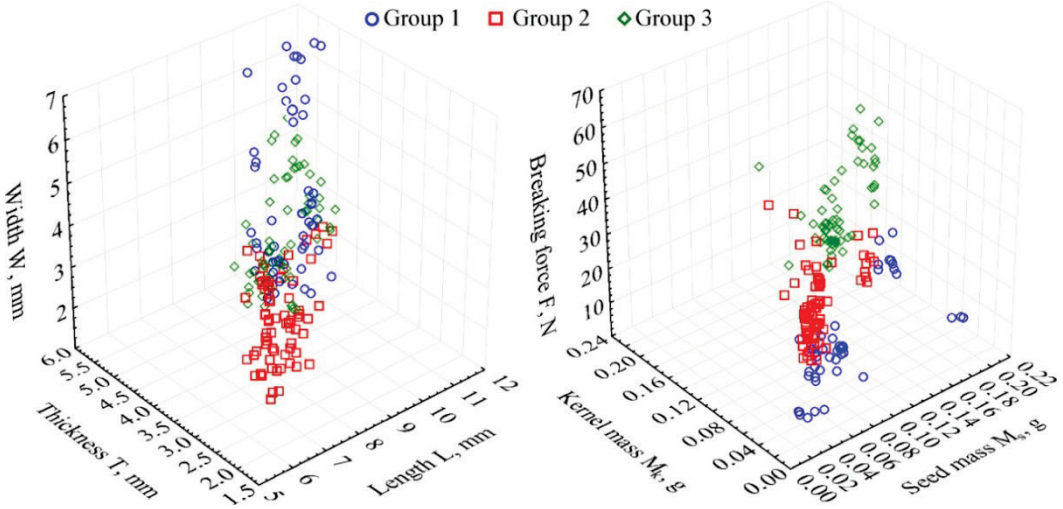

Fig. 5. Results of data clustering

In training the classification neural network, we selected the seed characteristics as the output data, the sample name as the target category, and the defined groups as the input values. For activation functions of neurons based on inputs and movement, the following options were considered: sigmoid function (Logistic) and hyperbolic tangent function (Tanh).

A multilayer perceptron (MLP) was chosen as the neural network – a modern artificial neural network with direct connections consisting of fully connected neurons with nonlinear activation functions, organized in at least three layers, notable for its ability to distinguish data that are not linearly separable (Guang-Bin et al., 2006; Mor et al., 2021). The number of hidden layers ranges from 5 to 15.

Table 4 shows that the most productive network is the MLP 30-15-3, which will be used as the base network. For the MLP 30-15-3 network, the weight coefficients and prediction table for data processing were calculated. The performance of the neural network was evaluated using data from other measurements of the same sunflower samples grown in nurseries. The testing performance of the created neural network was 95.6%, while the validation performance decreased to 96.7%. Some sample overlaps were observed, such as ZL169A (Group II) – ZL54A (Group II), ZL169A (Group II) – 168a (Group II), KP11A (Group I) – 165ar2 (Group I), SX75A (Group III) – 162bp1 (Group III), 178a (Group II) – 168b (Group II). However, within the adopted grouping, these samples remained within their respective groups.

The high performance quality of the neural network, as shown in Table 4, along with its successful validation on other data related to the research object, supports the development of a methodology. This methodology aims to systematize sunflower genotypes based on seed phenotypic characteristics using the data vector quantization method with neural networks (Fig. 6).

Table 5
Indicators of the computed neural networks

Network name	MLP 30-12-3	MLP 30-10-3	MLP 30-15-3	MLP 30-5-3	MLP 30-15-3
Training performance, %	93.93	93.18	99.72	98.48	99.35
Testing performance, %	97.66	96.78	99.43	97.54	98.61
Validation performance, %	96.42	98.54	99.67	98.25	97.39
Training algorithm	BFGS 9	BFGS 11	BFGS 10	BFGS 10	BFGS 12
Error function	SOS	SOS	SOS	Entropy	SOS
Activation function of hidden neurons	Tanh	Logistic	Logistic	Tanh	Logistic
Activation function of output neurons	Tanh	Tanh	Logistic	Softmax	Tanh

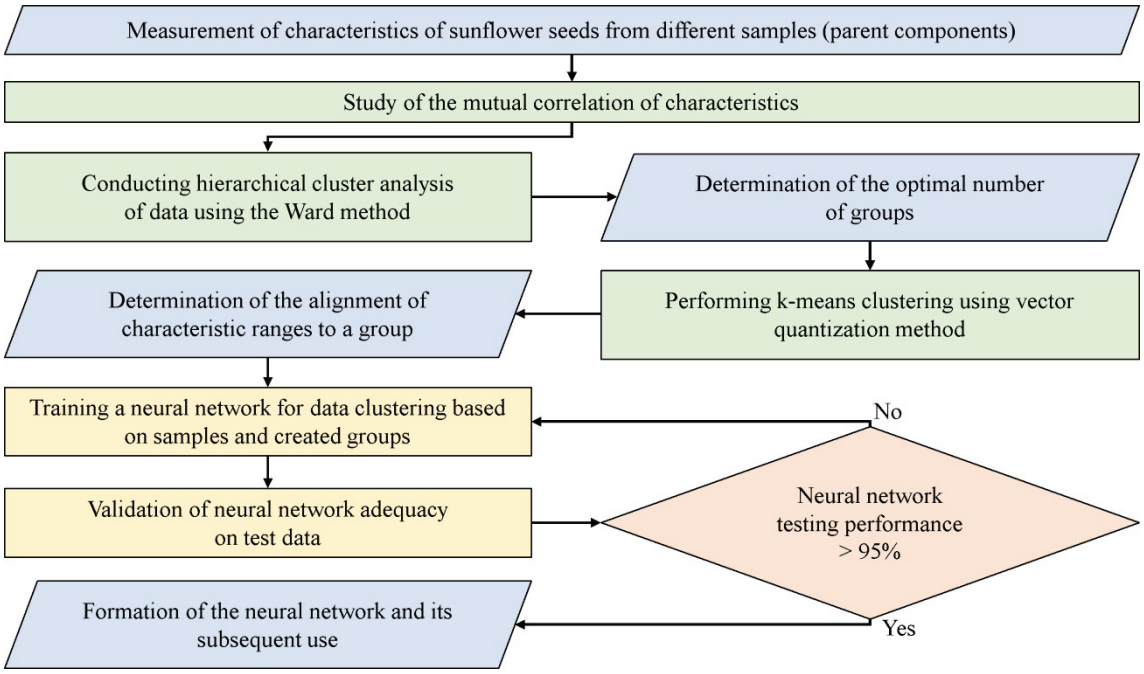

Fig. 6. Methodology for systematizing sunflower genotypes based on seed phenotypic characteristics using the data vector quan-tization method with neural networks

Discussion

In the study by Rajalakshmi et al. (2022) on sunflower seed classification, a deep neural network called RiceSeedNet was developed, which enabled the classification of rice seeds. The experimental results showed that RiceSeedNet achieves high accuracy of 98–99% for seed dataset classification. Although this is generally lower than the results of our research, the comparison is not very accurate due to the different types of crops.

In some studies (Li et al., 2020; Jin et al., 2022; Kolhar & Jagtap, 2023) that used machine learning methods for seed classification, accuracy values exceeding those achieved in our research were reported. These higher accuracies may be associated with greater differentiation between the classes distinguished in their studies. This consideration is consistent, as some sunflower seed samples are sufficiently similar in their morphological characteristics, making their classification significantly more challenging. Additionally, the strength of seed coat cracking, which is a crucial characteristic for confectionery sunflower use, was not considered in the cited studies.

The Crop-GPA platform (Gao et al., 2024) includes extensive data on phenotypic characteristics of various crops, and our research complements it both in terms of defined characteristics and samples from the Institute of Oilseed Crops of NAAS.

From the perspective of selective breeding, the best visual distribution (Fig. 7) was observed for seed coat cracking strength and seed mass. The highlighted first and third groups mainly contained large-seeded samples. In the first group, there were single-head (KP11A, KP11, 162B, 168B) and very large multi-head (340B, 165AR2) varieties. The third group comprised the varieties (Zaporizhian Confectionery, Prometey, Bilochka) and single-head lines (ZL50A, SK50A, 160B, 162BP1, 165VR1), which can also be considered quite large-seeded and require less seed hull cracking force. This distribution mostly confirms the selection conclusions made by other methods. Previous studies identified the best lines for large-seed characteristics as follows: single-head – KP11A and 160B, and multi-head 168B (Nosal et al., 2017; Vedmedeva & Nosal, 2020). During the tests of hybrids in two zones, the best parental component for large seeds was the line 340B, and the best hybrid was SK75A × 340B (Nosal et al., 2018). Comparing these results, it can be said that identifying the first and third groups is quite useful for confectionery sunflower breeders.

One limitation of our research was the use of samples from the Institute of Oilseed Crops of NAAS. Although other studies also used their own varieties and hybrids (Jin et al., 2022; Barrio-Conde et al., 2023), there is room for improvement by including additional samples.

Regarding our research, a relatively small group of lines was considered, used as parental components for oilseed and confectionery hybrids, including three varieties and several experimental lines created for confectionery purposes. This does not encompass the entire possible diversity of sunflower, excluding samples with low economic value and those significantly differing in seed qualities from the studied samples, which complicated the class differentiation. Furthermore, comparing classifier performance with varying numbers of classes showed difficulties when adding new varieties due to their similarity to one of the existing classes.

The equipment developed as part of the research (Fig. 1, 2) and the methodology for systematizing sunflower genotypes based on seed phenotypic characteristics (Fig. 6) can be applied in breeding for rapid analysis and screening of seed samples for further crossing.

Conclusion

The conducted research revealed the phenotypic characteristics of sunflower seeds (length, width, thickness, seed mass, kernel mass, and seed coat cracking strength) from different genotypes of sunflower bred by the Institute of Oilseed Crops of NAAS, grown in the southern Steppe of Ukraine. To achieve this, appropriate laboratory equipment

was developed: two modules for determining the morphological and rheological properties of the seeds. Moderate (according to the Cheddock scale) mutual correlations were identified for the geometric dimensions and mass of sunflower seeds from different samples (correlation coefficient of 0.6355–0.7528 and 0.5580–0.6918, respectively), and the lack of correlation was found for seed coat cracking strength and other characteristics (correlation coefficient – 0.0005–0.0179).

Based on the Ward's method and vector quantization, k-means clustering of the obtained research data was performed. As a result, sunflower samples were classified into three groups, which is also supported by previous research.

The methodology for systematizing sunflower genotypes based on seed phenotypic characteristics was developed using the vector quantization method with a neural network. The MLP neural network with a 30-15-3 architecture was developed for data grouping by samples and created groups using the Statistica software package. The training algorithm used was BFGS 10, and the activation function for hidden and output neurons was Logistic. The training performance of the developed neural network was 99.4%, testing performance was 95.6%, and validation performance was 96.7%.

The research was conducted as part of the scientific and technical work titled "Quantitative Phenotyping of Sunflower Genotypes" which is funded by an external European Union assistance instrument to fulfill Ukraine's commitments under the European Union Framework Programme for Research and Innovation "Horizon 2020".

References

Aliiev, E. B. (2020). Automatic phenotyping test of sunflower seeds. Helia, 43(72), 51–66.

Aliiev, E. B. (2023). The prospects of quantitative phenotyping of oilseed crops. Agrology, 6(3), 49–59.

Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. Information Sciences, 324, 126–145.

Barrio-Conde, M., Zanella, M. A., Aguiar-Perez, J. M., Ruiz-Gonzalez, R., & Gomez-Gil, J. (2023). A deep learning image system for classifying high oleic sunflower seed varieties. Sensors, 23(5), 2471.

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., & Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. Plant Genome, 14(3), e20112.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer, New York.

Brusco, M., Steinley, D., & Watts, A. L. (2024). Improving the Walktrap algorithm using k-means clustering. Multivariate Behavioral Research, 59(2), 266–288.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. Trends in Plant Science, 22(11), 961–975.

de Amorim, R. C. (2015). Feature relevance in ward's hierarchical clustering using the Lp norm. Journal of Classification, 32(1), 46–62.

Eldem, A. (2020). An application of deep neural network for classification of wheat seeds. Avrupa Bilim Ve Teknoloji Dergisi, 19, 213–220.

Gao, Y., Zhou, Q., Luo, J., Xia, C., Zhang, Y., & Yue, Z. (2024). CropGPA: An integrated platform of crop gene-phenotype associations. npj Systems Biology and Applications, 10, 15.

Guang-Bin, H., Qin-Yu, Z., & Chee-Kheong, S. (2006). Extreme learning machine: Theory and applications. Neurocomputing, 70(1), 489–501.

Hamerly, G., & Elkan, C. (2004). Learning the k in k-means. Advances in Neural Information Processing Systems, 16, 281.

Ibrar, D., Khan, S., Mahmood, T., Bakhsh, A., Aziz, I., Rais, A., Ahmad, R., Bashir, S., Nawaz, M., Rashid, N., Irshad, S., Alotaibi, S. S., Dvorackova, H., Dvoracek, J., & Hasnain, Z. (2022). Molecular markers-based DNA fingerprinting coupled with morphological diversity analysis for prediction of heterotic grouping in sunflower (*Helianthus annuus* L.). Frontiers in Plant Science, 13, 916845.

Ibrar, D., Khan, S., Raza, M., Nawaz, M., Hasnain, Z., Kashif, M., Rais, A., Gul, S., Ahmad, R., & Gaafar, A.-R. Z. (2024). Application of machine learning for identification of heterotic groups in sunflower through combined approach of phenotyping, genotyping and protein profiling. Scientific Reports, 14, 7333.

Jin, X., Zhao, Y., Wu, H., & Sun, T. (2022). Sunflower seeds classification based on sparse convolutional neural networks in multi-objective scene. Scientific Reports, 12, 19890.

Kolhar, S., & Jagtap, J. (2023). Plant trait estimation and classification studies in plant phenotyping using machine vision – A review. Information Processing in Agriculture, 10(1), 114–135.

Li, Z., Guo, R., Li, M., Chen, Y., & Li, G. (2020). A review of computer vision technologies for plant phenotyping. Computers and Electronics in Agriculture, 176, 105672.

Libbrecht, M., & Noble, W. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16, 321–332.

Luan, Z., Li, C., Ding, S., Wei, M., & Yang, Y. (2020). Sunflower seed sorting based on convolutional neural network. Eleventh International Conference on Graphics and Image Processing, 11373, 1K-1.

Mor, G., Roei, S., Jonathan, B., & Omer, L. (2021). Transformer feed-forward layers are key-value memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 5484–5495.

Najafabadi, Y. M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: Predicting yield from hyperspectral reflectance in soybean. Frontiers in Plant Science, 11, 624273.

Nosal O. O., Vedmedeva, K. V., Maklyak, K. M., & Leonova, N. M. (2018). Hospodars'ka otsinka krupnoplidnykh hibrydiv sonyashnyku [Economic assessment of large-seed sunflower hybrids]. Scientific and Technical Bulletin of the Institute of Oilseed Crops NAAS, 25, 83–95 (in Ukrainian).

Nosal, O. O., Vedmedeva, K. V., & Shkolova, S. V. (2017). Donors'ki vlastyvosti liniy sonyashnyku za oznakoyu krupnoplidnosti [Donor properties of sunflower lines based on large seeds]. Scientific and Technical Bulletin of the Institute of Oilseed Crops NAAS, 24, 110–121 (in Ukrainian).

Pasi, F. (2018). Efficiency of random swap clustering. Journal of Big Data, 5(1), 1–21.

Perez-Sanz, F., Navarro, P. J., & Egea-Cortines, M. (2017). Plant phenomics: An overview of image acquisition technologies and image data analysis algorithms. GigaScience, 6(11), gix092.

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

Van-Dijk, A. D. J., Kootstra, G., Kruijer, W., & Ridder, D. (2021). Machine learning in plant science and plant breeding. iScience, 24, 101890.

Vedmedeva, K. V., & Nosal, O. O. (2020). Otsinka krupnoplidnykh liniy sonyashnyku za kil'kisnymy kharakterystykamy morfolohichnykh oznak [Evaluation of large-seed sunflower lines by quantitative characteristics of morphological features]. Scientific and Technical Bulletin of the Institute of Oilseed Crops NAAS, 29, 46–55 (in Ukrainian).

Weylandt, M., Nagorski, J., & Allen, G. I. (2019). Dynamic visualization and fast computation for convex clustering via algorithmic regularization. Journal of Computational and Graphical Statistics, 29(1), 87–96.