

МАТЕМАТИЧНА МОДЕЛЬ СИСТЕМАТИЗАЦІЇ СЕЛЕКЦІЙНО-НАСІННИЦЬКОГО МАТЕРІАЛУ ОЛІЙНИХ КУЛЬТУР НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

Е. Б. Алієв¹, О. Ю. Алієва²

¹ Дніпровський державний аграрно-економічний університет

² Інститут олійних культур Національної академії аграрних наук України

У роботі запропоновано математичну модель систематизації генотипів соняшнику (*Helianthus annuus* L.) на основі фенотипових характеристик насіння, що дозволяє здійснювати ідентифікацію генотипів з урахуванням генетичних та екологічних факторів. Розроблена модель ґрунтується на модульному представленні фенотипових параметрів (довжина, товщина, ширина, маса насіння, маса ядра, сила руйнування) та їхньому зв'язку з дією зовнішніх факторів (елементи живлення, освітлення, температура, вологість, агротехніка). Математичний опис моделі реалізовано у вигляді тензорних рівнянь, що враховують варіативність фізіолого-генетичних систем і відхилення ознак у популяції. На основі експериментальних даних проведено кореляційний, кластерний і нейромережевий аналіз. Застосовано метод векторного квантування та багатосаровий перцептрон (MLP 30-15-3) з функцією активації Logistic і тренувальним алгоритмом BFGS 10. Висока ефективність розробленої моделі підтверджена точністю навчання (99,35%), тестування (95,63%) і валідації (96,67%). Отримані результати дозволяють здійснювати надійне групування соняшникових зразків, прогнозувати селекційні результати та оцінювати внесок генетико-фізіологічних систем у формування продуктивних ознак.

Ключові слова: соняшник, фенотипування, генотип, селекція, машинне навчання, багатосаровий перцептрон, систематизація, тензорна модель, кластеризація.

Вступ. В останні роки зростає доступність різноманітних наборів даних щодо фенотипування генотипів рослин (Najafabadi et al. 2021; Bayer et al. 2021; Van-Dijk et al. 2021). Однак основним викликом для сучасної науки про рослини є ефективне і точне прогнозування фенотипів на основі генотипів в умовах змінного клімату. Генетичні варіації впливають на біохімічні процеси в клітинах і тканинах, які в свою чергу, разом з екологічними сигналами, визначають розвиток органів, ріст рослин, врожайність та стійкість до абіотичних і біотичних факторів. У контексті сучасної селекції рослин дослідження впливу екологічних та генотипових змін відкриває нові горизонти в регуляції ключових процесів життєвого циклу рослин. Водночас, це також пов'язано з викликами щодо точності прогнозування врожайності сільськогосподарських культур в умовах змінного клімату.

Аналіз фенотипів на різних рівнях організації та розробка зв'язків між фенотипами та генотипами потребують інтеграції та обробки великих та різномірних наборів даних із вираженими шумами. Машинне навчання пропонує широкий спектр методів і підходів для виявлення прогностичних закономірностей у таких даних. Програми класифікації на основі машинного



*Corresponding author

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0

(<https://creativecommons.org/licenses/by/4.0/>)

навчання стають потужними інструментами для створення точних і надійних моделей, які використовуються для оцінки якості сільськогосподарської продукції. До цих програм входять різні алгоритми, такі як дерева рішень, штучні нейронні мережі, генетичні алгоритми, регресія та нечітка логіка. Ці алгоритми корисні для розробки моделей машинного навчання, які допомагають у виборі важливих ознак і корекції складних підходів до відображення вводу-виводу (Crossa et al. 2017; Libbrecht et al. 2015; Perez-Sanz et al. 2017).

Соняшник (*Helianthus annuus* L.) відомий своїм проявом гетерозису за багатьма ознаками, що робить його цікавим об'єктом дослідження з точки зору адаптивної інтрогресії та еволюційної біології (Adeleke et al. 2020; Heuzé et al. 2018). Соняшник цінується як модельна рослина для дослідження процесів розвитку квітів, а також як джерело харчової олії преміум-класу. Наразі соняшник є четвертою за значимістю олійною культурою у світі, завдяки своїй здатності рости в різних умовах і високій врожайності. Гібриди F1, отримані від віддалено споріднених високопродуктивних ліній, часто перевершують своїх батьків у продуктивності (Ibrar et al. 2022; Ibrar et al. 2024).

Машинне навчання вже відіграє значну роль у різних інженерних галузях, але його потенціал у прикладній селекції рослин, зокрема соняшнику, ще недостатньо досліджений. Зважаючи на це, постає необхідність у створенні ефективних математичних моделей, здатних підтримувати селекційні рішення на основі даних.

Математична модель генотипів соняшнику, що базується на переліку фенотипових характеристик та умов вирощування, повинна вирішувати наступні задачі: оцінка механізмів скринінгу батьківських пар для забезпечення заданого результату схрещування; оцінка вкладу генетичних особливостей батьківських пар; прогнозування селекційно-генетичних ознак у потомства; схрещування та отримання популяції F2; ідентифікація генотипів за їх фенотипами (Aliiev 2022; Aliiev et al. 2024).

Ці задачі взаємопов'язані і аспектом узагальненої проблеми – керування послідовними етапами селекційно-генетичного процесу.

Головною задачею є ідентифікація генотипів за фенотипами, так як вона дозволяє формулювати і вирішувати інші задачі. Її мета полягає в пошуку (або створенні) генотипу, що включає максимальну кількість позитивних зрушень заданих селекційних ознак. Тому в основу алгоритму цього завдання покладено принцип фонових ознак та принцип різноспрямованості зрушень кількісної ознаки окремої рослини під впливом генетичних та екологічних причин у багатовимірних ознакових координатах (тензорів).

Матеріали і методи досліджень. Розглянемо концептуальну схему класифікації генотипів за фенотипами, яка наведена на рис. 1.

Змінюючи класичну модель Фішера Р. запропонована математична модель генотипу соняшника, що описує інтегральну властивість продуктивності і-ї рослини:

$$\Psi_i = \mu + \gamma_{attr,i} + \gamma_{mic,i} + \gamma_{ad,i} + \gamma_{imm,i} + \gamma_{ef,i} + \gamma_{tol,i} + \gamma_{ont,i} + \gamma_{com,i} + \pi_{com,i} + \pi_{ont,i} + \pi_i, \quad (1)$$

де Ψ_i – фенотипове значення ознаки продуктивності у і-ї рослини; μ – середнє значення продуктивності у популяції; $\gamma_{attr,i}$ – відхилення атракції продуктів фотосинтезу зі стебел і листя в кошик; $\gamma_{mic,i}$ – відхилення розподілу продуктів атракції між насінням та ядром; $\gamma_{ad,i}$ – відхилення впливу системи адаптивності на продуктивність, що вимірюється загальною сухою біомасою рослини; $\gamma_{imm,i}$ –

вплив горизонтальної стійкості на продуктивність; $\gamma_{ef,i}$ – вплив оплати біомасою лімітуючих факторів ґрунтового харчування; $\gamma_{tol,i}$ – відхилення толерантності до загущення; $\gamma_{ont,i}$ – відхилення генетичної варіабельності тривалості фаз онтогенезу; $\gamma_{com,i}$ – відхилення генетичної конкуренції рослин за вологу, харчування, світло тощо; $\pi_{com,i}$ – відхилення негенетичної конкуренції, викликане неоднаковими початковими умовами зростання, $\pi_{ont,i}$ – відхилення, викликане зміною лімітуючих факторів в онтогенезі в період закладки і розвитку ознаки, π_i – середнє відхилення.

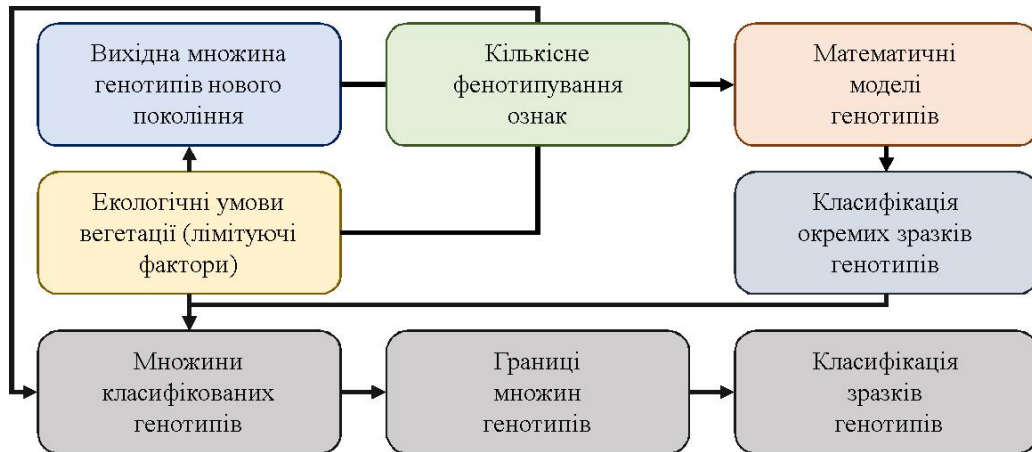


Рис. 1. Концептуальна схема класифікації генотипів за фенотипами

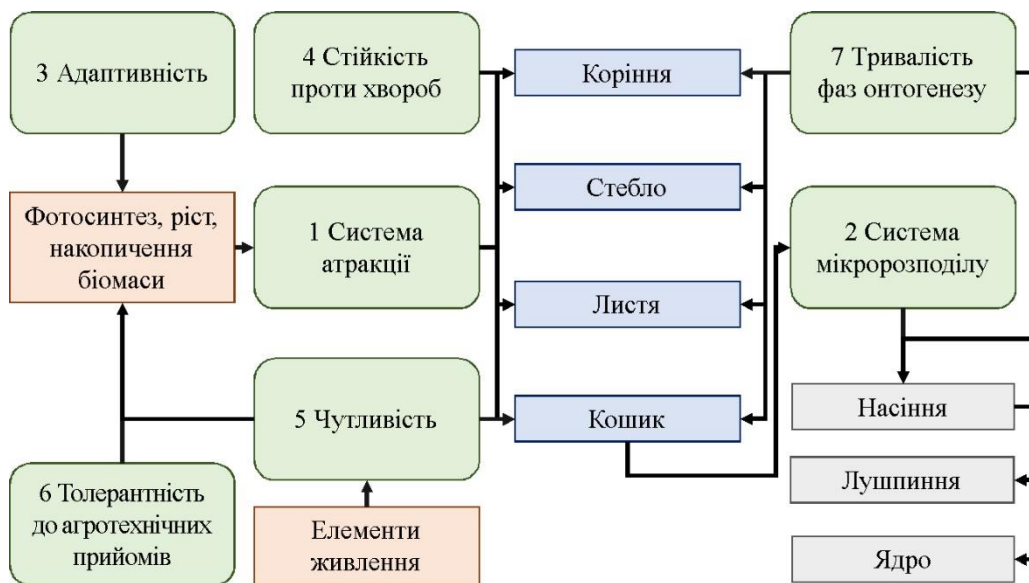


Рис. 2. Схема моделі генотипів соняшнику

Розшифровуємо кожну складову моделі (рис. 2) у вигляді конкретної співіснуючої генетично-фізіологічної системи і моделі модульної структури: 1 – система атракції (фенотипові параметри коріння φ_{root} , стебла φ_{stem} , листя φ_{leaf} і

кошика (квітки) φ_{head} ; 2 – система мікророзподілу (фенотипові параметри насіння φ_{seed} , лушпиння φ_{husk} і ядра φ_{kemel}); 3 – система адаптивності, то є стійкість до кліматичних і хімічних стресів середовища (ступінь затримання ростових процесів під впливом негативних факторів середовища, швидкість і час відновлення нормальних ростових процесів); 4 – система полігонного імунітету (стійкості рослин до шкідників і збудників хвороб, вироблення рослинами захисних речовин і механізмів); 5 – система сприйнятливості елементів живлення (параметри чутливості продуктивних характеристик до доз елементів живлення); 6 – система толерантності до агротехнічних прийомів; 7 – система варіабельності періодів онтогенезу (використовується в селекції з метою виведення критичної фази онтогенезу з-під впливу стресового фактору середовища).

Відобразимо наведені характеристики генетико-фізіологічних систем спочатку на структурі модульної моделі для соняшника (рис. 2):

$$\begin{aligned} \dot{\mathbf{X}}_i = \begin{pmatrix} \dot{\mathbf{x}}_{\text{root}} \\ \dot{\mathbf{x}}_{\text{stem}} \\ \dot{\mathbf{x}}_{\text{leaf}} \\ \dot{\mathbf{x}}_{\text{head}} \\ \dot{\mathbf{x}}_{\text{seed}} \\ \dot{\mathbf{x}}_{\text{husk}} \\ \dot{\mathbf{x}}_{\text{kemel}} \end{pmatrix}_i = \begin{pmatrix} \varphi_{\text{root},1} & \dots & \varphi_{\text{root},j} & \dots \\ \varphi_{\text{stem},1} & \dots & \varphi_{\text{stem},j} & \dots \\ \varphi_{\text{leaf},1} & \dots & \varphi_{\text{leaf},j} & \dots \\ \varphi_{\text{head},1} & \dots & \varphi_{\text{head},j} & \dots \\ \varphi_{\text{seed},1} & \dots & \varphi_{\text{seed},j} & \dots \\ \varphi_{\text{husk},1} & \dots & \varphi_{\text{husk},j} & \dots \\ \varphi_{\text{kemel},1} & \dots & \varphi_{\text{kemel},j} & \dots \end{pmatrix}_i \left[\begin{pmatrix} A_{f,1} & F_{\text{fert},1}(t) \\ \dots & \dots \\ A_{f,j} & F_{\text{fert},j}(t) \\ \dots & \dots \end{pmatrix} + \begin{pmatrix} A_{l,1} & F_{\text{light},1}(t) \\ \dots & \dots \\ A_{l,j} & F_{\text{light},j}(t) \\ \dots & \dots \end{pmatrix} \right] + \\ + \begin{pmatrix} A_{t,1} & F_{\text{temp},1}(t) \\ \dots & \dots \\ A_{t,j} & F_{\text{temp},j}(t) \\ \dots & \dots \end{pmatrix} + \begin{pmatrix} A_{m,1} & F_{\text{moist},1}(t) \\ \dots & \dots \\ A_{m,j} & F_{\text{moist},j}(t) \\ \dots & \dots \end{pmatrix} + \begin{pmatrix} A_{a,1} & F_{\text{agro},1}(t) \\ \dots & \dots \\ A_{a,j} & F_{\text{agro},j}(t) \\ \dots & \dots \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_j \\ \dots \end{pmatrix} \end{aligned} \quad (2)$$

де $\dot{\mathbf{X}}_i$ – функція зміни параметрів рослини соняшнику; φ_{root} , φ_{stem} , φ_{leaf} , φ_{head} , φ_{seed} , φ_{husk} , φ_{kemel} – фенотипові параметри коріння, стебла, листя, кошика (квітки), насіння, лушпиння і ядра відповідно; j – поточний номер фенотипового параметра; F_{fert} , F_{light} , F_{temp} , F_{moist} , F_{agro} – динамічні функції впливу елементів живлення, світлового фактору, температурного фактору, водозабезпечення та агротехнічних прийомів відповідно; A_f , A_l , A_t , A_m , A_a – параметри моделі; ε – випадкові обурення, що відображають інформаційну невизначеність моделі.

Представимо модель (2) у більш компактній тензорній формі:

$$\dot{\mathbf{X}}_i = \varphi_i \left[A_f F_{\text{fert}}(t) + A_l F_{\text{light}}(t) + A_t F_{\text{temp}}(t) + A_m F_{\text{moist}}(t) + A_a F_{\text{agro}}(t) \right] + \varepsilon \quad (3)$$

в якій всі змінні та параметри об'єднані у відповідні тензори.

Модель (3) визначає стан i -ї рослини, при цьому вплив лімітуючих факторів неоднаковий для всіх рослин, як і дія генетико-фізіологічних систем, що призводять до обурення станів окремих рослин і виникнення середових та генетичних дисперсій. Такі обурення можна представити у такому вигляді:

$$\begin{aligned} \Delta \mathbf{X}_i = \mathbf{X} - \mathbf{X}_i = \mathbf{U}_E \Delta \mathbf{E}_i + \mathbf{U}_\varphi \Delta \varphi_i, \\ \mathbf{U}_E = \frac{\partial \mathbf{X}_i}{\partial \mathbf{E}}, \quad \mathbf{U}_\varphi = \frac{\partial \mathbf{X}_i}{\partial \varphi}, \end{aligned} \quad (4)$$

де U_E , U_φ – вектори функцій чутливості стану модуля відповідно до екологічних і генетичних збурень; ΔE_i , $\Delta \varphi_i$ – вектори варіацій спостерігаємих екологічних факторів і неспостерігаємих генетичних впливів.

Вираз (4) відображає моделювання вклади екологічних та генетичних факторів. Разом з тим селекціонер найчастіше має справу з варіаціями ознак, які будемо позначати як ΔY_i . У цьому випадку сенс класифікації генотипів полягає в встановленні причин варіацій ознак окремих рослин, що спостерігаються, в порівнянні з середньопопуляційними значеннями. У разі, якщо такими причинами служать екологічні чинники, маємо справу з модифікаціями одного й того ж генотипу, а при встановленні генетичних причин – з новим генотипом.

Введемо на розгляд квадратичний функціонал якості класифікації:

$$J_i = \int_{t_0}^t \left[(\Delta X_i(t) - \Delta Y_i(t))^T (\Delta X_i(t) - \Delta Y_i(t)) \right] dt \quad (5)$$

форма якого відображає «розбалансування» модельованих і спостережуваних зрушень селекційних ознак, що виникають за рахунок не спостережуваних впливів семи генетико-фізіологічних систем.

Мінімізація критерію (2.5) по не спостережуваним впливам генетичних факторів $\Delta \varphi_i$ для відомих варіацій екологічних факторів $\Delta E_i(t)$ дозволить оцінити вклади кожної з генетико-фізіологічних систем у прояв змін у аналізованій рослині:

$$\Delta \hat{\varphi}_i = \arg \min_{\Delta \varphi_i} \rightarrow \int_{t_0}^t \left[(\Delta X_i(t | \Delta E_i) - \Delta Y_i(t))^T (\Delta X_i(t | \Delta E_i) - \Delta Y_i(t)) \right] dt \quad (6)$$

Тепер, маючи області допустимих значень впливів генетико-фізіологічних систем для окремих генотипів Ω_k , де $k = 1, 2, 3, \dots, K$ – індекси генотипів (класів), можна визначити вирішальне правило класифікації:

$$k_i = k^*, \text{ якщо } \Delta \varphi_i \in \Omega_k^*. \quad (7)$$

Маючи в своєму розпорядженні вихідну інформацію про відхилення екологічних факторів $\Delta E_i(t)$ та спостережувані відхилення кількісних ознак і-ї рослини $\Delta Y_i(t)$, за допомогою рівнянь (2)–(7) може визначити її приналежність до генотипу із заданими ознаками. Фіксуючи під час проведення цієї процедури для кожного класу множини $\Delta E_k(t)$, $\Delta Y_k(t)$, можна побудувати межі між генотипами у просторі екологічних відхилень та варіацій кількісних ознак у вигляді спеціальної апроксимуючої функції:

$$L_k = \Phi_{k,k+1}(\Delta E, \Delta Y). \quad (8)$$

У цьому випадку вирішальне правило виглядає так:

$$k_i = k^*, \text{ якщо } \Phi_{k,k+1}(\Delta E, \Delta Y) - c \leq 0. \quad (9)$$

де c – граничне значення правила, що є одним з параметрів вирішального правила.

Таким чином, розглянута вся схема ідентифікації генотипів за фенотипами, що спостерігаються. Тут вся процедура попередньої сепарації генотипів і модифікацій осіб становить етап навчання з більш простого вирішального правила (8)–(9).

Результати досліджень та їхнє обговорення. Наведена математична модель (2)–(9) є відносно класичною і не враховує сучасні підходи зокрема

застосування нейронних мереж. Тому спираючись на отримані дані щодо морфологічний і реологічних властивостей насіння соняшнику проведемо їх обробку.

У результаті проведення обмірювання насінин різних зразків основного поля отримано набір вихідних даних у вигляді таблиці середніх значень їх основних характеристик (рис. 3). Обробку отриманих даних проведемо в декілька етапів (Aliyev et al, 2024; Vedmedeva et al, 2023; Vedmedieva, 2023).

1	2	3	4	5	6	7
Назва	Довжина L, мм	Товщина T, мм	Ширина W, мм	Маса насіння Ms, г	Маса ядра Mk, г	Сила руйнування F, Н
1 ЗЛ50А	7,65	3,42	3,8	0,07845	0,05783	58
2 ЗЛ50А	7,55	3,27	3,67	0,07307	0,05365	59,5
3 ЗЛ50А	7,51	3,16	3,49	0,06972	0,05177	54,5
4 ЗЛ50А	7,59	3,38	3,89	0,06896	0,04549	50,1
5 ЗЛ50А	7,28	3,13	3,18	0,06983	0,03629	48,2
6 ЗЛ50А	7,17	2,4	3,67	0,07008	0,02963	49,4
7 ЗЛ50А	7,47	3,43	3,5	0,07189	0,05441	49,5
8 ЗЛ50А	7,32	3,26	3,69	0,06767	0,04523	52,1
9 ЗЛ50А	7,02	3,32	3,21	0,06823	0,03766	53
10 ЗЛ169А	7,84	3,39	3,32	0,05559	0,05355	41,1
182 178а x 238	6,38	2,41	3,29	0,04094	0,03542	31,1
183 165Вр1	7,27	2,51	3,63	0,07875	0,04681	45,9
184 165Вр1	7,29	2,5	3,65	0,07813	0,04643	45,3
185 165Вр1	7,38	2,48	3,53	0,07756	0,04655	44,6
186 КП11	10,43	5,84	4,04	0,17676	0,09797	18,7
187 КП11	10,27	5,69	3,82	0,17693	0,10243	19,1
188 КП11	10,29	5,69	3,94	0,17748	0,10069	19,1

Рис. 3. Вихідні дані основних характеристик насіння соняшнику різних зразків основного поля

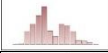
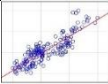
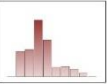
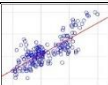
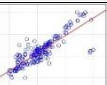
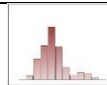
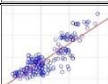
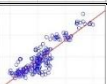
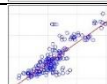
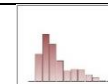
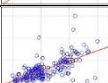
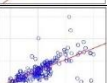
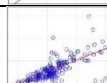
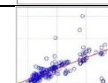
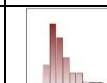
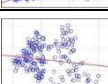
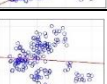
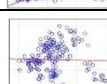
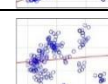
Першим етапом проведемо дослідження взаємної кореляції характеристик насіння соняшнику різних зразків (табл. 1). На графіках вісь абсцис відповідає назвам колонки таблиці, вісь ординат відповідає назвам строк таблиці, сині точки – експериментальні дані, червона пряма – лінійна пряма регресії між двома характеристиками насіння; на гістограмах представлено розподіл характеристик насіння, яка відповідає колонці і рядку таблиці; рівень статистичної значимості $P < 0.05$; $t^t(0.05; 187) = 1.973$ – табличне значення t-критерія Стюдента.

Спираючись на методі Уорда на другому етапі проведемо ієрархічний кластерний аналіз даних для визначенні кількості груп, на які можна розбити дані. Для візуалізації результатів ієрархічної кластеризації даних побудуємо дендрограму (рис. 4).

Вона показує ступінь близькості окремих об'єктів і скупчень, а також наочно демонструє в графічній формі послідовність їх об'єднання або поділу. Кількість рівнів дендрограми відповідає числу кроків злиття або розщеплення кластерів. А кількість розгалужень відповідає можливій кількості груп. Бачимо, що при відстані зв'язків 170–240 розгалуження знаходяться на одному рівні, тому кількість груп приймаємо рівною 3.

Таблиця 1

Графічна візуалізація і розраховані коефіцієнти кореляції (r) і t -критерій Стьюдента (t_r) взаємної кореляції характеристик насіння соняшнику різних зразків

Характеристика	Довжина L , мм	Товщина T , мм	Ширина W , мм	Маса насіння M_s , г	Маса ядра M_k , г	Сила руйнування F , Н
Довжина L , мм		$r = +0,7528$ $t_r = 9,124 > t_r^t$	$r = +0,6355$ $t_r = 5,921 > t_r^t$	$r = +0,7070$ $t_r = 8,035 > t_r^t$	$r = +0,4593$ $t_r = 7,646 > t_r^t$	$r = -0,0179$ $t_r = 0,755 < t_r^t$
Товщина T , мм			$r = +0,7147$ $t_r = 5,002 > t_r^t$	$r = +0,7768$ $t_r = 8,079 > t_r^t$	$r = +0,5060$ $t_r = 7,393 > t_r^t$	$r = -0,0045$ $t_r = 0,114 < t_r^t$
Ширина W , мм				$r = +0,6918$ $t_r = 5,771 > t_r^t$	$r = +0,4922$ $t_r = 9,132 > t_r^t$	$r = -0,0005$ $t_r = 0,587 < t_r^t$
Маса насіння M_s , г					$r = +0,5580$ $t_r = 10,82 > t_r^t$	$r = +0,0102$ $t_r = 0,781 < t_r^t$
Маса ядра M_k , г						$r = +0,0336$ $t_r = 0,210 < t_r^t$
Сила руйнування F , Н						

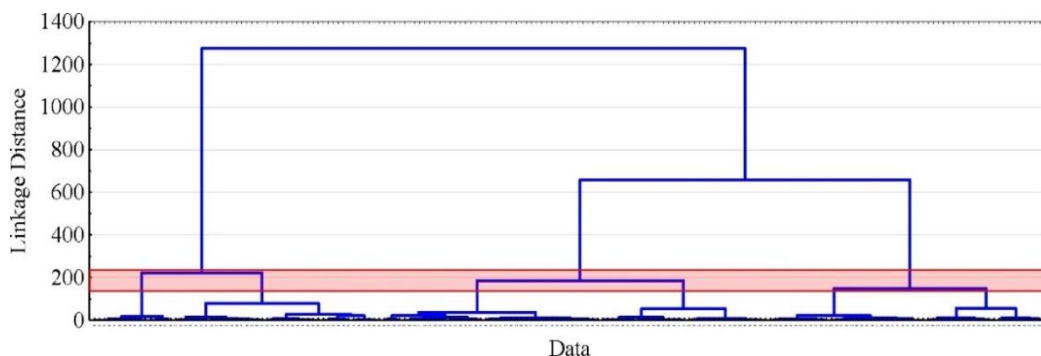


Рис. 4. Дендрограма відстаней між кроками результатів ієрархічної кластеризації даних

Далі на третьому етапі провівши кластеризацію k -середніх, тобто методом векторного квантування, який має на меті розбити n спостережень на k груп, в яких кожне спостереження належить до кластера з найближчим середнім значенням – центри кластера або центроїд кластера. Провівши ранжування середніх для кожної групи в межах від 0 до 1 (де 0 – найменше значення, 1 – найбільше значення) побудовано графічну візуалізацію отриманого групування (рис. 5).

Результати групування даних наведено на рисунку 6. Тривимірні залежності наочно демонструють розділення зразків на три групи. Так до 1 групи відносяться наступні зразки: КП11А, КП11, 340В, 162В, 165ар2, 164г, 168в. До 2 групи відносяться: ЗЛ169А, ЗЛ54А, 178а, 168б, 174д, 175бп1, 174е, 168б, 168а, 178а × 238. До 3 групи відносяться: ЗЛ50А, СХ75А, 160в, 162бп1, Білочка, Прометей, Запорізький кондитерський, 165Вр1.

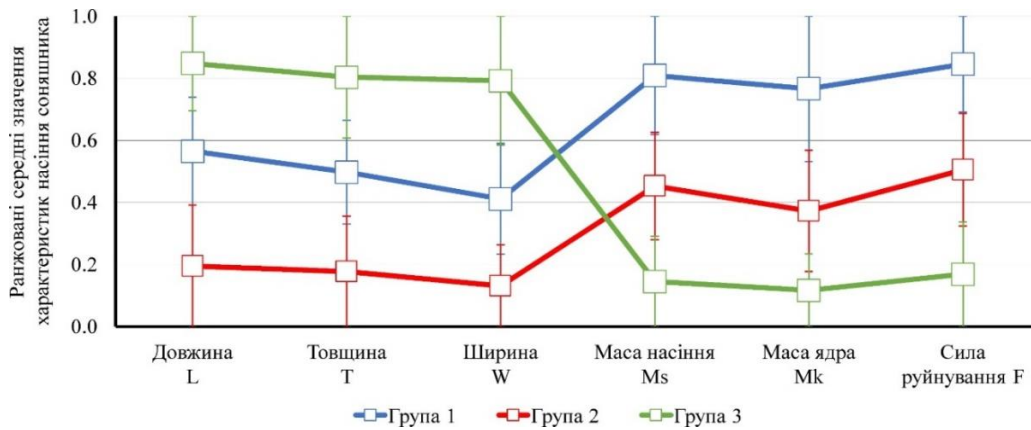


Рис. 5. Графік ранжованих середніх значень груп для кожної характеристики насіння соняшника

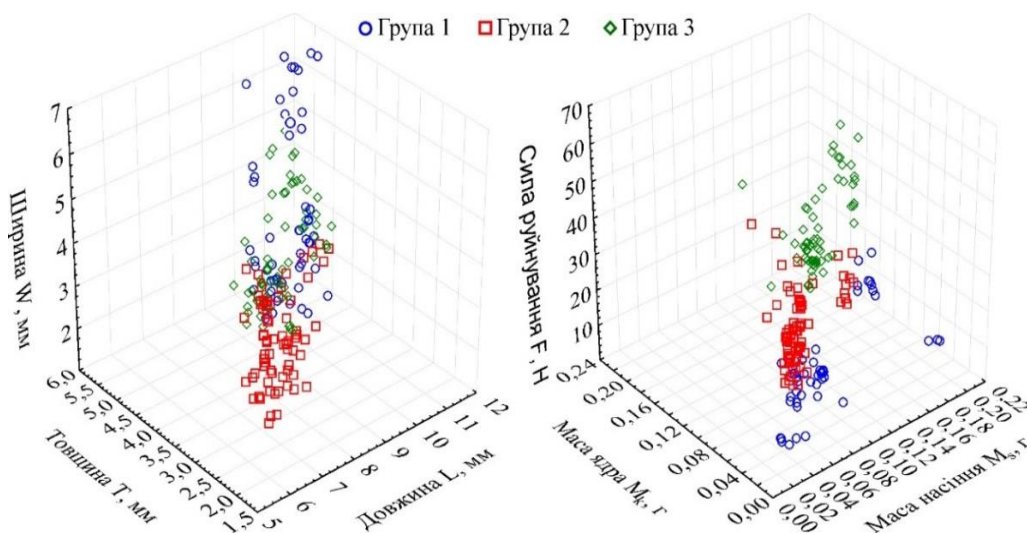


Рис. 6. Результати групування даних

Четвертим етапом є навчання нейронної мережі (Neural Networks) групування даних по зразках і створених групах. Нейронна мережа складається зі з'єднаних блоків або вузлів, які називаються штучними нейронами, які вільно моделюють нейрони мозку. Вони з'єднані ребрами, які моделюють синапси головного мозку. Кожен штучний нейрон отримує сигнали від підключених нейронів, потім обробляє їх і посилає сигнал іншим підключеним нейронам.

Сигнал є дійсним числом, а вихід кожного нейрона обчислюється деякою нелінійною функцією суми його входів, яка називається функцією активації. Сила сигналу при кожному з'єднанні визначається вагою, яка регулюється в процесі навчання.

В якості нейронної мережі обрано багатошаровий персептрон (MLP) – сучасна штучна нейронна мережа прямого зв'язку, що складається з повністю з'єднаних нейронів з нелінійною функцією активації, організованих принаймні в три шари, примітних здатністю розрізняти дані, які не є лінійно відокремлюваними. Кількість прихованих блоків – від 5 до 15.

Встановлено, що найбільш продуктивною є мережа MLP 30-15-3, яку і будемо використовувати в якості базової. Для мережі MLP 30-15-3 розраховані вагові коефіцієнти і таблиця прогнозів при обробці даних. Якість роботи нейронної мережі оцінено використовуючи данні іншого вимірювання тих самих зразків соняшнику, які вирощені в розплідниках. Продуктивність тестування для створеної нейронної мережі склала 95,63%, а продуктивність валідації знизилась до 96,67%. При цьому спостерігалось пересікання деяких зразків з іншими ЗЛ169А (ІІ група) – ЗЛ54А(ІІ група), ЗЛ169А (ІІ група) – 168а (ІІ група), КП11А (І група) – 165ар2 (І група), СХ75А (ІІІ група) – 162бп1 (ІІІ група), 178а (ІІ група) – 168б (ІІ група). Однак в рамках прийнятого групування зазначені зразки знаходяться в межах своїх груп (рис. 7).

Network weights			Predictions spreadsheet for Групи			
Weight ID	Connections	Weight values	Samples: Train, Test, Validation			
	3.MLP 30-15-3	3.MLP 30-15-3	Case name	Sample	Групи Target	Групи - Output 3. MLP 30-15-3
1	Довжина, мм --> hidden neuron 1	-9,1777	1	Train	3	3
2	Товщина, мм --> hidden neuron 1	-9,5953	2	Train	3	3
3	Ширина, мм --> hidden neuron 1	-7,0410	3	Validation	3	3
4	Маса насіння, г --> hidden neuron 1	-5,9988	4	Train	3	3
5	Маса ядра, г --> hidden neuron 1	-1,4689	5	Train	3	3
6	Сила руйнування, Н --> hidden neuron 1	13,1992	6	Train	3	3
7	Назва(160в) --> hidden neuron 1	0,4517	7	Train	3	3
8	Назва(162В) --> hidden neuron 1	-7,0040	8	Test	3	3
9	Назва(162бп1) --> hidden neuron 1	0,1760	9	Train	3	3
10	Назва(164г) --> hidden neuron 1	-5,7911	10	Train	2	2
11	Назва(165Вр1) --> hidden neuron 1	0,1882	11	Train	2	2
12	Назва(165ар2) --> hidden neuron 1	-6,9689	12	Train	2	2
13	Назва(168а) --> hidden neuron 1	4,2024	175	Train	2	2
500	hidden neuron 5 --> Групи(3)	-1,7390	176	Test	2	2
501	hidden neuron 6 --> Групи(3)	24,6180	177	Test	2	2
502	hidden neuron 7 --> Групи(3)	-8,1400	178	Train	2	2
503	hidden neuron 8 --> Групи(3)	15,6569	179	Train	2	2
504	hidden neuron 9 --> Групи(3)	-7,1156	180	Validation	2	2
505	hidden neuron 10 --> Групи(3)	-2,3205	181	Validation	2	2
506	hidden neuron 11 --> Групи(3)	-2,9896	182	Train	2	2
507	hidden neuron 12 --> Групи(3)	-2,7819	183	Train	3	3
508	hidden neuron 13 --> Групи(3)	12,2296	184	Train	3	3
509	hidden neuron 14 --> Групи(3)	-17,3152	185	Train	3	3
510	hidden neuron 15 --> Групи(3)	7,6010	186	Train	1	1
511	hidden bias --> Групи(1)	3,6879	187	Train	1	1
512	hidden bias --> Групи(2)	-4,6098	188	Validation	1	1
513	hidden bias --> Групи(3)	-3,8386				

а

б

Рис. 7. Розраховані вагові коефіцієнти мережі MLP 30–15–3 (а) і таблиця прогнозів при обробці даних (б)

Достатньо висока якість роботи нейронної мережі та її перевірка на інших даних у рамках об'єкта дослідження дає змогу скласти методикку систематизації генотипів соняшнику за фенотиповими характеристиками насіння на основі методу векторного квантування даних з використанням нейронної мережі (рис. 8).

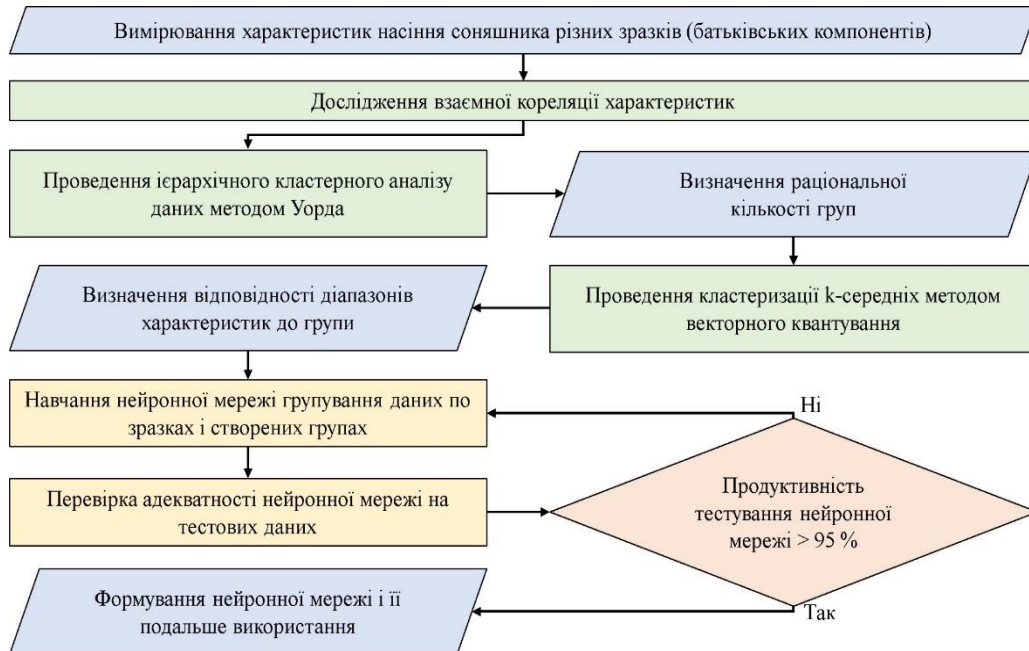


Рис. 8. Методика систематизації генотипів соняшнику за фенотиповими характеристиками насіння на основі методу векторного квантування даних з використанням нейронної мережі

Висновки. Запропонована формалізована математична модель ідентифікації генотипів соняшнику за їх фенотипами, що включає в себе оцінювання величин ненаглядних вкладів генетико-фізіологічних систем у продуктивність рослини. Складена схема моделі генотипів соняшнику і проведено її математичний опис у тензорному вигляді: зміна параметрів рослини соняшнику залежить від фенотипових параметрів коріння φ_{root} , стебла φ_{stem} , листя φ_{leaf} , кошика (квітки) φ_{head} , насіння φ_{seed} , лушпиння φ_{husk} і ядра φ_{kemel} при динамічному впливі елементів живлення $F_{\text{fert}}(t)$, світлового фактору $F_{\text{light}}(t)$, температурного фактору $F_{\text{temp}}(t)$, водозабезпечення $F_{\text{moist}}(t)$ та агротехнічних прийомів $F_{\text{agro}}(t)$.

Розроблена методика систематизації генотипів соняшнику за фенотиповими характеристиками насіння на основі методу векторного квантування даних з використанням нейронної мережі. Розроблена нейронна мережа MLP 30-15-3 для групування даних по зразках і створених групах в програмному пакеті Statistica. При цьому алгоритм тренування – BFGS 10, а функція активації прихованих і вихідних нейронів – Logistic. Продуктивність навчання для створеної нейронної мережі склали 99,35%, продуктивність тестування – 95,63%, а продуктивність валідації – 96,67%.

References

- Adeleke BS, Babalola OO (2020) Oilseed crop sunflower (*Helianthus annuus*) as a source of food: Nutritional and health benefits. *Food Science & Nutrition*. 9. 4666–4684 DOI: 10.1002/fsn3.1783
- Aliiev EB (2022) Automatic Phenotyping of Sunflower Seed Material: Monograph. Kyiv: Agrarian Science. 104 p.
- Aliiev EB, Vedmedieva KV (2024) Quantitative Phenotyping of Sunflower Genotypes: Monograph. Dnipro: LIRA. 204 p.
- Aliyev E, Vedmedeva K, Machova T, Vedmedev S (2024) Study of the distribution of phenotypic characteristics of sunflower seeds in a head of different genotypes. *Bulgarian Journal of Crop Science*. 61(4): 73–89 DOI: 10.61308/JIXX8922
- Bayer PE, Petereit J, Danilevich MF, Anderson R, Batley J, Edwards D (2021) The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome*. 14(3). e20112 DOI: 10.1002/tpg2.20112
- Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burguño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*. 22(11): 961–975 DOI: 10.1016/j.tplants.2017.08.011
- Heuzé V, Tran G, Hassoun P, Lessire M, Lebas F (2018) Sunflower hulls and sunflower screenings. Feedipedia, a programme by INRAE, CIRAD, AFZ and FAO. <https://feedipedia.org/node/733>
- Ibrar D, Khan S, Mahmood T, Bakhsh A, Aziz I, Rais A, Ahmad R, Bashir S, Nawaz M, Rashid N, Irshad S, Alotaibi SS, 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 DOI: 10.3389/fpls.2022.916845
- Ibrar D, Khan S, Raza M, Nawaz M, Hasnain Z, Kashif M, Rais A, Gul S, Ahmad R, Gaafar ARZ (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 DOI: 10.1038/s41598-024-58049-z
- Libbrecht M, Noble W (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. 16: 321–332 DOI: 10.1038/nrg3920
- Najafabadi YM, Earl HJ, 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. DOI: 10.3389/fpls.2020.624273
- Perez-Sanz F, Navarro PJ, Egea-Cortines M (2017) Plant phenomics: An overview of image acquisition technologies and image data analysis algorithms. *GigaScience*. 6: 1–18 DOI: 10.1093/gigascience/gix092
- Van-Dijk ADJ, Kootstra G, Kruijer W, Ridder D (2021) Machine learning in plant science and plant breeding. *iScience*. 24. 101890. DOI: 10.1016/j.isci.2020.101890
- Vedmedieva KV (2023) Marker Morphological Traits of Sunflower: Identification, Inheritance, and Allelic Gene Effects: Monograph. Kyiv: Agrarian Science. 240 p.

Vedmedeva K, Nosal O, Poliakova I, Machova T (2023) Correlations of confectionary seed traits in different head zones sunflower. *Helia*. 46(79): 215–231
DOI: 10.1515/helia-2023-0012

MATHEMATICAL MODEL FOR THE SYSTEMATIZATION OF BREEDING AND SEED MATERIAL OF OIL CROPS BASED ON NEURAL NETWORKS

E. B. Aliiev¹, O. Yu. Aliieva²

¹ Dnipro State Agrarian and Economic University

² Institute of Oilseed Crops of the National Academy of Agrarian Sciences of Ukraine

This study presents a formalized mathematical model for the systematization of sunflower (*Helianthus annuus* L.) genotypes based on the phenotypic characteristics of seeds. The research addresses urgent challenges in modern agricultural science, particularly the need for accurate phenotype prediction from genotypic data under changing climate conditions. The proposed model aims to serve as a tool for effective decision-making in plant breeding and contributes to the implementation of the "digital breeding" concept, which integrates morphological, biometric, physico-mechanical, and genetic traits. The model captures the complex interplay between internal (genetic) and external (environmental) factors affecting the expression of yield-related traits in sunflower. It is based on a system of differentiated phenotypic parameters such as seed length, thickness, width, total seed weight, kernel weight, and rupture force, all of which are linked to distinct genetic-physiological systems (attraction system, microdistribution, adaptability, tolerance, immune response, etc.). The mathematical description is formulated using tensor equations, enabling the modeling of plant state changes as a result of multifactorial influences. The practical part of the study involves a multi-stage data analysis pipeline. Initially, correlation analysis was conducted to examine relationships between seed traits. Then, hierarchical clustering using Ward's method and k-means vector quantization were applied to classify seed samples into three distinct groups with high intra-group consistency. To automate genotype classification, a multilayer perceptron (MLP) neural network with a 30-15-3 architecture was trained. The training used the BFGS algorithm (Broyden–Fletcher–Goldfarb–Shanno) and logistic activation functions, resulting in high performance: training accuracy – 99.35%, testing – 95.63%, validation – 96.67%. Based on the high accuracy of the developed neural network model, a methodology for sunflower genotype systematization was developed, using vector quantization and machine learning. This approach enables early screening of parental lines, prediction of trait inheritance in offspring, assessment of genetic similarity, and identification of novel genotypes. Importantly, the proposed model can be adapted to other oilseed and agricultural crops by incorporating species-specific morphometric traits. In summary, this study marks a significant advancement toward the automation of plant breeding processes, the development of high-yield sunflower hybrids, and the integration of neural networks into bioinformatics and agricultural technology. The results provide a foundation for building intelligent decision-support systems for breeding practice.

Key words: sunflower, phenotyping, genotype, plant breeding, neural network, multilayer perceptron, tensor model, clustering, vector quantization, breeding automation.