Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 14(4), 1448-1460, 2024
Journal of the Institute of Science and Technology, 14(4), 1448-1460, 2024

ISSN: 2146-0574, eISSN: 2536-4618

Biosystem Engineering **DOI: 10.21597/jist.1526542**

Research Article

Received: 08.08.2024 **Accepted: 31.08.2024**

To Cite: Altikat, A. & Alma, M.H. (2024). Advanced Predictive Analytics in Agriculture: Case Study on Wheat Kernel Weight. *Journal of the Institute of Science and Technology*, 14(4), 1448-1460.

Advanced Predictive Analytics in Agriculture: Case Study on Wheat Kernel Weight

Alperay ALTIKAT[1*], Mehmet Hakkı ALMA[2]

Highlights:

- ANN models performed best in predicting thousand-grain weight of 13 wheat varieties, with an R^2 value of 0.866.
- Increasing input variables improved MLR accuracy but risked overfitting. ANN outperformed MLR models significantly.
- The PCA+MLR model was ineffective, with low R^2 values (0.24-0.31). The PCA+ANN model greatly improved accuracy, achieving an R^2 of 0.981.
- ANN and PCA+ANN models provide high prediction accuracy. MLR models offer moderate prediction capabilities.

Keywords:

- Thousand-kernel weight
- Wheat
- Multiple linear regression
- Artificial neural networks
- Principal component analysis
- Hybrid modeling

ABSTRACT:

This research, was aimed at modeling the thousand-grain weight of 13 different wheat varieties using five different input parameters. We used multiple linear regression (MLR), artificial neural networks (ANN), principal component analysis (PCA), and two different hybrid models consisting of PCA + MLR and PCA + ANN for this purpose. The MLR models were tested with various input configurations, demonstrating moderate explanatory power, with R^2 values ranging from 0.37 to 0.44. Increasing the number of independent variables increased prediction accuracy but also increased the risk of overlearning. ANN models showed significantly higher performance in prediction accuracy. The best performance was achieved in the ANN20 architecture with an R2 value of 0.866. In this architecture, a combination of the gradient descent training function, the hyperbolic tangent sigmoid transfer function, the linear transfer function, and 18 neurons were used. The PCA+MLR hybrid model was not effective in predicting thousand-grain weight. The fact that R^2 values obtained with different input configurations vary between 0.24 and 0.31 shows that the prediction accuracy of the model is low. In contrast, the PCA+ANN hybrid model significantly improved the prediction accuracy, and the best model achieved an R^2 value of 0.981, an RMSE of 0.0829, and an MAE of 0.0359. The PCA+ANN model, which preserved the necessary variance by reducing the complexity of the input data, enabled the ANN to focus on the most critical components for accurate prediction. This study demonstrates that whereas ANN and PCA+ANN models give significantly increased accuracy in predicting wheat varieties' thousand-kernel weights, MLR models only offer moderate prediction capabilities.

[1] Alperay ALTIKAT (**Orcid ID:**0009-0005-8270-1728), Mehmet Hakkı ALMA (**Orcid ID:** 0000-0001-6323-7230), Iğdır University Agriculture Faculty Department of the Biosystems Engineering, Iğdır, Türkiye

***Corresponding Author:** Alperay ALTIKAT, e-mail: alper.altikat@igdir.edu.tr

INTRODUCTION

Modeling seed physical qualities has major applications in a variety of sectors, including optimizing agricultural equipment design, assessing seed quality, and enhancing seed processing operations. By modeling the physical properties of seeds, time-consuming and costly analyzes are avoided. In this way, clearer and more practical information about seeds is obtained and the infrastructure for a quality production process is created (Gierz et al., 2022). For example, detailed information about the quality of the seed can be obtained with models with high accuracy. In particular, preliminary information is obtained about important factors for vegetative development such as the design of the discs used in the planter arrangements of pneumatic planting machines and the planting depth. With effective modeling, engineering designs of seed silos can be made with precision(Arigela et al., 2021).

At the same time, seed selection machines can be designed thanks to model studies. In this way, seeds of homogeneous sizes can be collected in the same class. This increases the performance of the planting machine at the time of planting and paves the way for homogeneous germination. Optimum results can be achieved by using appropriate models to adjust the airflow rate used at the time of planting, especially in pneumatic planting machines(Kaliniewicz et al., 2019).

Thousand-kernel weight is one of the important parameters showing the quality of the seed. Low thousand-kernel weight can easily deform under mechanical loads. In addition, the germination percentage of these seeds is lower. Predicting seed quality can be aided by modeling thousand-kernel weight. Research conducted based on storage circumstances revealed that seeds with a lower thousand-kernel weight had a fall in durability and germination rates while those with a higher weight showed more favorable indicators (Dryha et al., 2022; Polishchuk & Konovalov, 2023)

In a study examining the effects of thousand-kernel weight on plant development and seedling growth rate, it was concluded that the increase in thousand-kernel weight contributed positively to the early growth of seedlings (Thangjam & Sahoo, 2016). It is possible to find different studies in the literature to model thousand-kernel weight. In a study conducted to model the thousand-seed weight of the canola plant, factors such as different varieties, planting norms, and physiological quality criteria were used as input parameters. As a result of the research, the thousand-kernel weight of the canola plant was modeled with high accuracy (Ferreira et al., 2017).

Artificial neural networks are a frequently used method in modeling studies. Due to the non-linear relationships between production parameters, especially in agricultural production, the effectiveness of artificial neural networks is among the factors investigated in most studies in this field. For example, in a study, some environmental factors such as soil nitrogen level, air temperature, and precipitation rate were used as input into the growth period of wheat. As a result of the research, it was concluded that the most accurate results were obtained in models made with the artificial neural networks method (Mamann et al., 2019). Shamsabadi et al. (Shamsabadi et al., 2022) used the artificial neural networks method in a research conducted to model the yields of different wheat hybrid seeds. To design a phenotype identification system, Zhang et al.(Zhang et al., 2023) used seven different characteristics of wheat seed as input parameters and modeled thousand-kernel weight. Al-Adhaileh and Aldhyani(Al-Adhaileh & Aldhyani, 2022) used the artificial neural networks method in their research to model the yield of wheat and other seeds. As a result of the research, wheat yield was modeled with high accuracy. In a similar study, artificial neural networks and other machine-learning methods were used to predict the yield of different cannabis varieties. As a result of the research, hemp yield could be modeled with high accuracy (Sieracka et al., 2023). In another study by Saffariha et al.

(Saffariha et al., 2020) the germination rates of Salvia Limbata seeds were successfully modeled using the artificial neural network method. Fonseca de Oliveira et al. (Fonseca de Oliveira et al., 2022) modeled the quality parameters of peanut seeds using artificial intelligence methods.

In this research, which was conducted to model the thousand-kernel weight of seeds, hybrid models as well as artificial neural networks (ANN), multi-component analysis (PCA), and artificial neural networks (ANN) methods were used. PCA+ANN and PCA+MLR methods were used as a hybrid model in the research. For this purpose, five different input sets were used in all models and it was investigated which input set would produce the best model. 13 different wheat varieties were used in the research. The difference of this research from other studies is that it is not limited to just a single model or input set, but offers a broader perspective by comparing the performances of different combinations and hybrid methods. This increases the generalizability of research results and contributes to obtaining more accurate and reliable results in the modeling process.

MATERIALS AND METHODS

Thirteen different wheat varieties were used in the research (Figure 1). A digital caliper was used to measure the length (L), width (W) and thickness (T) of the seeds. Based on these values, arithmetic (D_a) and geometric (D_g) mean diameter values were calculated with the help of the equations specified in Equation 1 and Equation 2, respectively.

$$D_a = \frac{L+W+T}{3} \quad \text{(1)}$$

$$D_g = \sqrt[3]{L * W * T} \quad \text{(2)}$$

The research employed models in which the seed thousand kernel weights were estimated using five alternative combinations of variety, width, length, thickness, arithmetic mean diameter, and geometric mean diameter values as input factors (Table 1).

Table 1. Input-output parameters used in all models in the research

Input no	Input	Output
1	Variety + width	Thousand-kernel weight
2	Variety + width + length	
3	Variety + width + length + thickness	
4	Variety + width + length + thickness + AMD	
5	Variety + width + length + thickness + AMD + GMD	

AMD: Arithmetic mean diameter, GMD: Geometric mean diameter

Nacibey Karma Müfitbey Çetinel

Figure 1. Wheat varieties analyzed in the study

The Modelling with Multiple Linear Regression

Equation 3 describes the Multiple Linear Regression (MLR) approach, whereas Figure 2 depicts the model architecture. In the equation, Y is the projected value of the model, x is input ai, i=0 to n, is the regression coefficient.

$$Y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \quad \text{.....(3)}$$

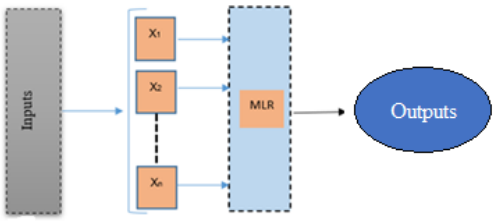

Figure 2. The architecture of the MLR Model

To reduce the number of input parameters, principal component analysis (PCA) was used. These new input parameters were called principal components (PC-eigenvectors), and MathWorks MATLAB was used to construct them. By default, MATLAB's PCA function uses the singular value decomposition (SVD) algorithm and returns the percentage of the total variance explained by each principal component.

Artificial Neural Network (ANN) Based Modeling

An artificial neural network (ANN) is another model utilized in the study. Artificial neural networks are widely employed in modeling studies involving variables, particularly those with non-linear associations. Using the right amount of neurons, transfer and activation functions, and learning algorithms while taking the problem's structural requirements into account, models are created using this technique (Gardner & Dorling, 1998). In the study, four distinct neuron counts, three transfer functions, and two learning functions were combined to create ANN structures that modeled thousand-kernel weight (Table 2). Figure 3 shows the architecture of an artificial neural network.

Table 2. Modeling Techniques Utilizing Artificial Neural Networks (ANN)

Input	ANN Structures			Output parameter
	Learning functions	Transfer functions	Neurons	
Input no 1	Traingdm	T-T	2	Thousand-kernel weight
Input no 2			10	
Input no 3		T-P	18	
Input no 4	Traingd		26	
Input no 5		P-P		

Traingdm: Gradient descent with momentum training function, Traingd: Gradient descent training function T: Hyperbolic tangent sigmoid transfer function, P: Linear transfer function

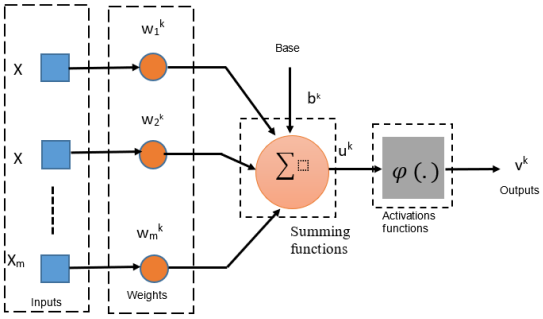

Figure 3. Artificial neural network architecture

Combining Multiple Linear Regression Models with Principal Component Analysis Integration

PCs were approved as input parameters in this technique for modeling thousand-kernel weight, and the MLR approach was integrated with PCs (Figure 4). The primary aspect of analysis yielded PCs.

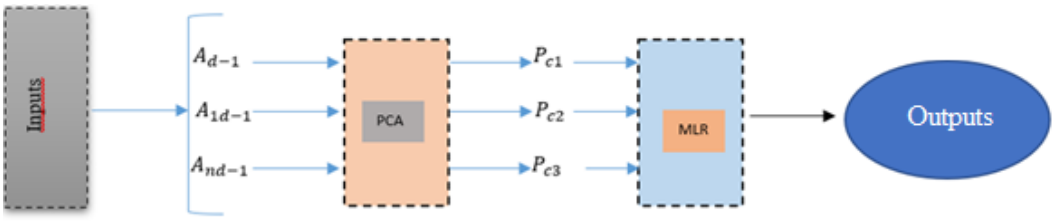

Figure 4. Structural Framework of PCA Integrated with MLR

Principal Component Analysis with Artificial Neural Network

The PCs were used as input parameters in this method as in the PCA + MLR method. The same transfer-learning functions and neuron numbers used in the ANN method were used together with PCs for modeling thousand-kernel weight. Figure 5 illustrates the architecture of principal component analysis with the artificial neural network.

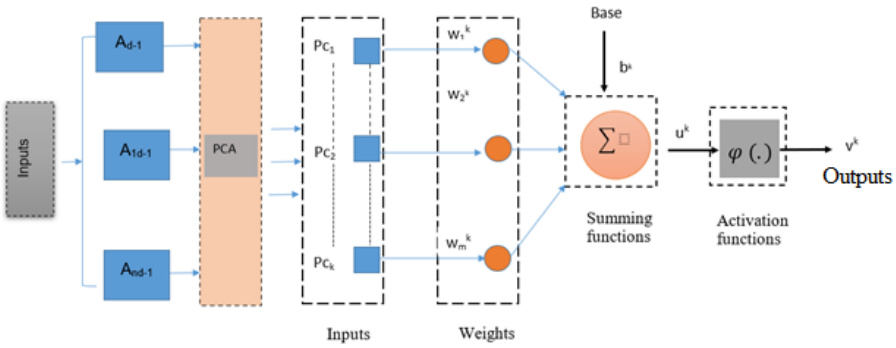

Figure 5. The architecture of PCA with ANN

In this study, the training set received 70% of the data, while the test and verification sets received 15% each. This data was partitioned for the ANN. The R values were examined both during performance verification and post-training to evaluate the networks' learning capacity. When a network's R-value got close to 1, it was considered to have been trained effectively. MATLAB software (R2019a), the most popular tool for predicting air pollution levels, was used to develop the ANN frameworks.

Performance Evaluation for Models

Model accuracy was verified using mean absolute error (MAE), R-squared (also called the coefficient of determination or R^2), and root mean-square error (also called root mean square deviation or RMSE). When RMSE and MAE approach 0 and R2 reaches 1, a model is considered very accurate.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{pi} - Y_{di}\right)^2} \quad \ldots\ldots(3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_{pi} - Y_{di}\right| \quad \ldots\ldots(4)$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}\left(Y_{pi}-Y_{di}\right)^2}{\sum_{i=1}^{n}\left(Y_{pi}-\bar{Y}\right)^2}\right) \quad \ldots\ldots(5)$$

In these equations; where, n is the number of observations, Y_{pi} is the predicted value for observation i, Y_{di} is the real value from observation i, and $\bar{Y}$ is the average of the real value.

RESULTS AND DISCUSSION

The Results of Multiple Linear Regression (MLR) Modeling

Statistical data of the multiple linear regression results performed to model thousand-kernel weight according to different input parameters are presented in Table 3. Accordingly, R^2 values for 5 different inputs vary between 0.37 and 0.437, and the accuracy of the model increased with increasing the number of inputs in MLR and the highest accuracy rate was obtained for input number 5. The effect of increasing the number of independent variables in multiple linear regression (MLR) on R^2 increases the explanatory power of the model but also brings the risk of overfitting (Uyanık & Güler, 2013). Additionally, the use of adjusted R^2 in cases where R^2 may be inflated is important to reduce the effects of overfitting (Mittlböck & Heinzl, 2002). Each new independent variable added explains some of the variance in the model, increasing the total variance. This increase is significant if the independent variables are truly related to the dependent variable (Schielzeth, 2010). However, adding too many independent variables may cause the model to overfit the learning set and lose its ability to generalize by learning the noise in the learning set (Ghasemzadeh et al., 2024).

Table 3. The statistical results and equations of MLR analysis

Input	Statistics			Equations
	R^2	F	P	
1	0.37	37.23	0.000	$y = 22.07 + 0.43X_1 + 4.47X_2$
2	0.41	29.37	0.000	$y = 14.5 + 0.41X_1 + 3.66X_2 + 1.57X_3$
3	0.43	23.19	0.000	$y = 12.81 + 0.384X_1 + 3.11X_2 + 1.2X_3 + 2.28X_4$
4	0.43	18.41	0.000	$y = 12.86 + 0.38X_1 - 2.36X_2 - 4.26X_3 - 3.18X_4 + 16.38X_5$
5	0.44	15.89	0.000	$y = 11.96 + 0.38X_1 - 19.4X_2 - 4.5X_3 - 24.1X_4 + 26.9X_5\ 35.3X_6$

X_1=Variety, X_2= Width, X_3= Length, X_4=Thickness, X_5= AMD, X_6=GMD

The Results of Artificial Neural Network

In the research, R2 values of five different input sets are presented in Table 4. The highest R^2 value (0.866) was obtained in the model using the first input set and the statistical results of the models adapted to different network architectures are shown in Table 5. The findings of this study reveal that the traingd learning function generally performs higher than the traingdm learning function. In particular, the ANN20 model (trainingd, T-P, with 18 neurons) performed best in all metrics. The R^2 value of this model was determined as 0.866. This result shows that the model explains 86.6% of the variance in the data set. RMSE and MAE values of the same model are 0.219 and 0.160, respectively, supporting the accuracy of predictions with low error rates. In terms of transfer functions, the combination of hyperbolic tangent sigmoid and linear transfer function (T-P) generally showed higher performance. Additionally, models with higher neuron counts (18 and 26 neurons) were found to perform better compared to models with lower neuron counts (2 and 10 neurons). This combination increases the prediction accuracy and strengthens the generalization ability of the model, especially when used with the traingd learning function. The regression analysis results and expected and observed values of the ANN20 architecture are shown in Figure 6, and the performance of this network architecture is shown in Figure 7.

Table 4. R^2 changes according to different inputs used in the ANN method

	Learning functions	Transfer functions	Neuron number	Input1	Input2	Input3	Input4	Input5
				R^2	R^2	R^2	R^2	R^2
ANN1	traingdm	T-T	2	0.168	0.377	0.099	0.450	0.444
ANN2	traingdm	T-P	2	0.012	0.253	0.026	0.002	0.063
ANN3	traingdm	P-P	2	0.370	0.408	0.366	0.350	0.349
ANN4	traingdm	T-T	10	0.084	0.566	0.561	0.383	0.543
ANN5	traingdm	T-P	10	0.001	0.093	0.470	0.033	0.568
ANN6	traingdm	P-P	10	0.003	0.032	0.412	0.181	0.425
ANN7	traingdm	T-T	18	0.127	0.514	0.548	0.188	0.611
ANN8	traingdm	T-P	18	0.714	0.028	0.563	0.547	0.052
ANN9	traingdm	P-P	18	0.310	0.021	0.365	0.036	0.139
ANN10	traingdm	T-T	26	0.113	0.259	0.271	0.372	0.286
ANN11	traingdm	T-P	26	0.781	0.035	0.527	0.133	0.341
ANN12	traingdm	P-P	26	0.287	0.026	0.346	0.029	0.263
ANN13	traingd	T-T	2	0.390	0.448	0.272	0.029	0.431
ANN14	traingd	T-P	2	0.303	0.357	0.436	0.444	0.421
ANN15	traingd	P-P	2	0.364	0.403	0.199	0.410	0.359
ANN16	traingd	T-T	10	0.662	0.136	0.551	0.431	0.584
ANN17	traingd	T-P	10	0.548	0.531	0.378	0.551	0.332
ANN18	traingd	P-P	10	0.368	0.355	0.424	0.416	0.407
ANN19	traingd	T-T	18	0.643	0.471	0.581	0.609	0.601
ANN20	**traingd**	**T-P**	**18**	**0.866**	0.457	0.548	0.511	0.518
ANN21	traingd	P-P	18	0.370	0.410	0.409	0.421	0.369
ANN22	traingd	T-T	26	0.787	0.002	0.624	0.574	0.469
ANN23	traingd	T-P	26	0.633	0.608	0.588	0.557	0.567
ANN24	traingd	P-P	26	0.364	0.411	0.396	0.294	0.417

Traingdm: Gradient descent with momentum training function, Traingd: Gradient descent training function T: Hyperbolic tangent sigmoid transfer function, P: Linear transfer function

Table 5. Statistical results of Input no 1

	Learning functions	Transfer functions	Neuron number	R^2	RMSE	MAE
ANN1	traingdm	T-T	2	0.168	0.545	0.436
ANN2	traingdm	T-P	2	0.012	0.618	0.528
ANN3	traingdm	P-P	2	0.370	0.481	0.389
ANN4	traingdm	T-T	10	0.084	0.590	0.457
ANN5	traingdm	T-P	10	0.001	0.863	0.665
ANN6	traingdm	P-P	10	0.003	0.635	0.534
ANN7	traingdm	T-T	18	0.127	0.595	0.428
ANN8	traingdm	T-P	18	0.714	0.321	0.264
ANN9	traingdm	P-P	18	0.310	0.554	0.453
ANN10	traingdm	T-T	26	0.113	0.604	0.509
ANN11	traingdm	T-P	26	0.781	0.280	0.211
ANN12	traingdm	P-P	26	0.287	0.523	0.408
ANN13	traingd	T-T	2	0.390	0.469	0.396
ANN14	traingd	T-P	2	0.303	0.499	0.381
ANN15	traingd	P-P	2	0.364	0.480	0.385
ANN16	traingd	T-T	10	0.662	0.349	0.278
ANN17	traingd	T-P	10	0.548	0.403	0.324
ANN18	traingd	P-P	10	0.368	0.475	0.386
ANN19	traingd	T-T	18	0.643	0.360	0.293
ANN20	**traingd**	**T-P**	**18**	**0.866**	**0.219**	**0.160**
ANN21	traingd	P-P	18	0.370	0.477	0.388
ANN22	traingd	T-T	26	0.787	0.278	0.193
ANN23	traingd	T-P	26	0.633	0.363	0.273
ANN24	traingd	P-P	26	0.364	0.477	0.385

Traingdm: Gradient descent with momentum training function, Traingd: Gradient descent training function T: Hyperbolic tangent sigmoid transfer function, P: Linear transfer function

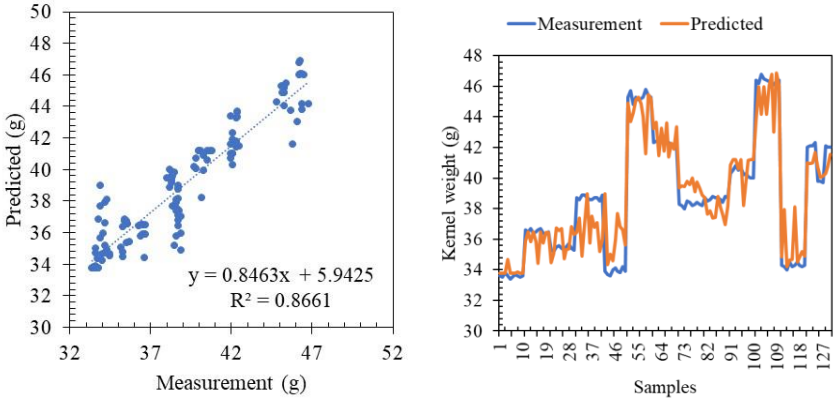

Figure 6. Regression analysis, expected and observed values for ANN 20

Recent research highlights the superior performance of the traingd learning function over the traingdm function in artificial neural networks (ANNs). A study demonstrated that models employing traingd generally outperformed those using traingdm in various metrics, particularly when high neuron counts were utilized. Additionally, other studies align with these findings, showing that simpler architectures like the two-dimensional Spiking Neuron Model, compared to more complex models, can yield better classification results due to their lower miss-classification rates (Kandpal & Mehta, 2019).

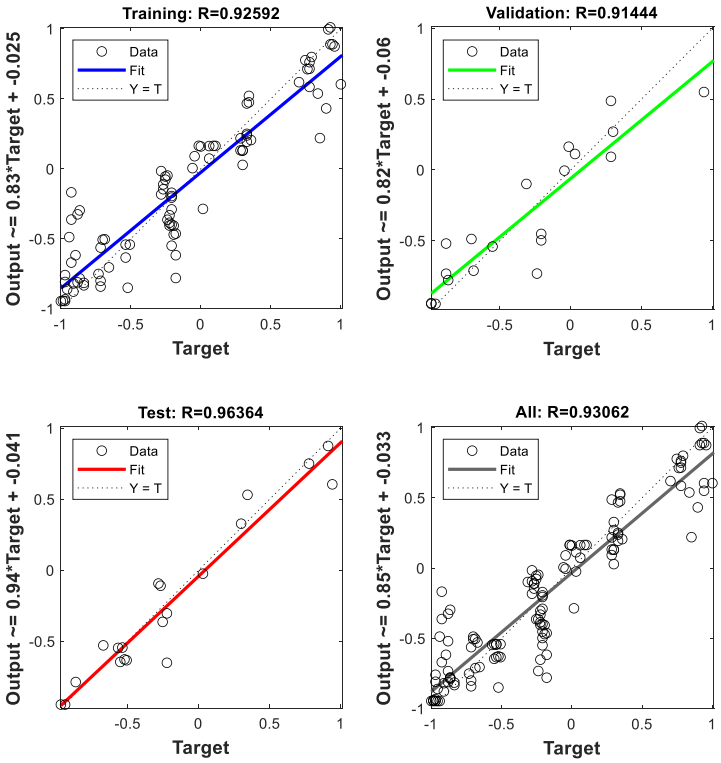

Figure 7. Network performance of ANN 20 architecture

In another study, it was stated that in modeling studies using artificial neural networks, better learning and accuracy rates were obtained in architectures where more neurons were used (Hankook et al., 1990). In another study investigating the effect of the number of neurons on model performance, it was stated that higher number of neurons exhibited better network performance. This increase in performance was caused by the strengthening of the generalization capacity of the network and the increase in recognition ability (Balda & Mathar, 2018).

The results of PCA+MLR

In the statistical results of the PCR+MLR hybrid model in the study, R^2 values were found to be very low in all input sets (Tablo 6). In this hybrid model, an R^2 value of 0.31 was reached in the set of 5 inputs with the highest number of inputs, but it was concluded that this value was insufficient to model the thousand grain weight of the seed.

Table 6. The statistical results and equations of PCA+MLR

Input	Statistics			Equations
	R^2	F	P	
1	0.24	40.93	0.000	$y = 37.63 + 2.34X_1 + 0X_2$
2	0.27	15.92	0.000	$y = 35.30 + 2.53X_1 - 2.38X_2 + 1.24X_3$
3	0.29	13.13	0.000	$y = 34.59 + 2.96X_1 + 2.41X_2 + 1.15X_3 - 4.70X_4$
4	0.29	10.46	0.000	y =34.5 +2.9X1 +37.6X2 +36.3X3 +30.3X4-105.3X5
5	0.31	9.49	0.000	y =33.9 +3X1 +19.4X2 +23.8X3 +11.0X4-98.2X5+42.6X6

X_1=Variety, X_2= Width, X_3= Length, X_4=Thickness, X_5= AMD, X_6=GMD

The result of PCA+ANN hybrid modeling for the prediction of kernel weight

In this hybrid model, seed thousand kernel weight was modeled by combining ANN20 architecture, which gives the best results in the ANN method, with PCA. The R^2 values in this hybrid model are significantly higher compared to the other models and the best result is found in the PCA+ANN model using the 1st input (Table 7).

Table 7. Statistical analysis results of the PCA+ANN model

Input no	Principal components					Statistical Results		
	PC1	PC2	PC3	PC4	PC5	R^2	RMSE	MAE
1	99.00	0.01	X	X		0.981	0.0829	0.0359
2	96.87	2.34	0.79	X	X	0.975	0.0949	0.0662
3	96.43	2.48	0.81	0.28	X	0.979	0.0872	0.0378
4	95.78	3.10	0.84	0.28	X	0.978	0.0878	0.0494
5	95.21	3.59	0.91	0.29	X	0.979	0.0867	0.0406

The model that used the first input had the best accuracy and the lowest error rates. These findings show that the PCA+ANN hybrid model, particularly in models where the initial principal components are mostly used, is capable of accurately predicting the thousand-kernel weight. The predicted and expected results from the regression analysis of the model using the 1st input are shown in Figure 8, and the network performances of this model are presented in Figure 9.

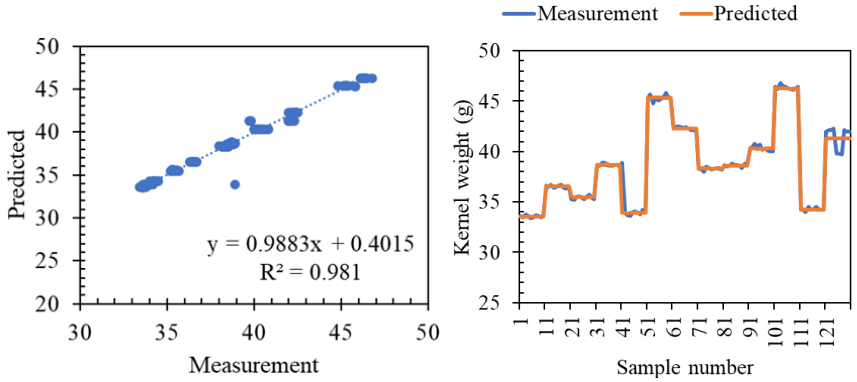

Figure 8. Expected and predicted values with regression analysis of the PCA+ANN model

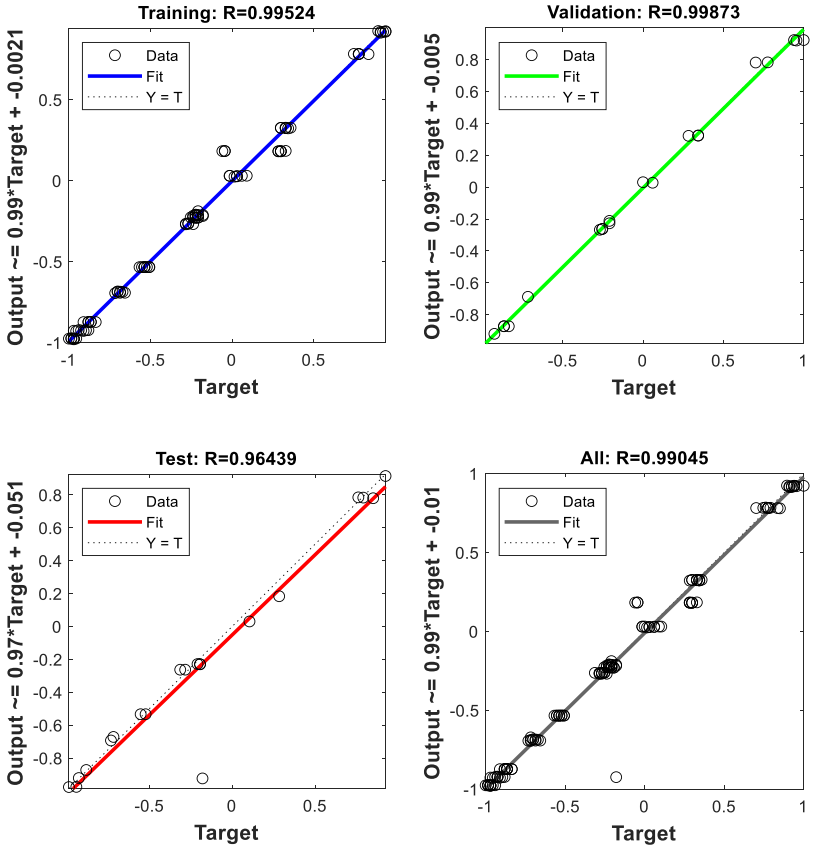

Figure 9. Network performance of the PCA+ANN model

Predictive modeling using Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) has proven to be highly effective in many agricultural applications. For example, Kheir et al., (Kheir et al., 2023) used ANN and hybrid models to model the thousand-kernel weight of wheat. As a result of the research, they declared that the best results were obtained in hybrid models.

CONCLUSION

The R^2 values in this hybrid model are significantly higher compared to the other models and the best result is found in the PCA+ANN model using the 1st input (Table 7).

Last but not least, the R^2 values of the MLR models range from 0.37 to 0.44 for various input setups, indicating a modest predictive power. While adding additional independent factors improved the accuracy of predictions, there was a chance of overlearning. With five input parameters, the greatest R^2 value was 0.44, suggesting that in complicated agricultural data sets, MLR might not be enough to produce high-accuracy predictions.

Predictive accuracy was significantly greater for the ANN models. The best performing ANN20 model reached an R^2 value of 0.866, explaining 86.6% of the variance in the data set. This model recorded low RMSE and MAE values using trainingd learning function, combination of tan-sigmoid and purelin transfer functions, and 18 neurons. Models with higher neuron counts generally performed better, indicating that network complexity is important in improving prediction accuracy.

The PCA+MLR hybrid model did not perform effectively in predicting thousand-kernel weight. R^2 values obtained with different input configurations varied between 0.24 and 0.31, indicating that the prediction accuracy of the model is low. The poor performance of the model highlights the difficulty of capturing complex and nonlinear relationships in agricultural data sets with linear models.

The PCA+ANN hybrid model significantly increased the prediction accuracy, and the best model achieved an R^2 value of 0.981, RMSE of 0.0829, and MAE of 0.0359. By reducing the complexity of the input data, the PCA+ANN model preserved the necessary variance and enabled the ANN to focus on the most critical components for accurate prediction. The highest accuracy achieved with the first set of inputs demonstrates that careful selection and preprocessing of inputs is critical to optimize model performance.

ACKNOWLEDGEMENTS

We would like to thank Iğdır University Scientific Research Projects Unit.

Conflict of Interest

There is no conflict of interest between

Author's Contributions

The authors declare that they have contributed equally to the article.

REFERENCES

Al-Adhaileh, M. H., & Aldhyani, T. H. H. (2022). Artificial intelligence framework for modeling and predicting crop yield to enhance food security in Saudi Arabia. *PeerJ Comput Sci*, *8*, e1104. https://doi.org/10.7717/peerj-cs.1104

Arigela, A., Kvs, R., & kumar, A. (2021). Study of Physical Properties of Zea mays in the Development of Seed Metering Unit. *International Journal of Agriculture Environment and Biotechnology*, *14*, 159-163. https://doi.org/10.30954/0974-1712.02.2021.5

Balda, E. B. A., & Mathar, R. (2018). An Information Theoretic View on Learning of Artificial Neural Networks. *IEEE International Conference on Signal Processing and Communication Systems*. https://doi.org/10.1109/ICSPCS.2018.8631758

Dryha, V. V., Doronin, V. A., Kravchenko, Y. A., Doronin, V., & Orlov, S. D. (2022). The effect of the storage conditions on the quality of switchgrass seeds of different 1000-kernel weight. *Scientific Papers of the Institute of Bioenergy Crops and Sugar Beet*.

Ferreira, A. S., Zucareli, C., Junior, A. A. B., Werner, F., & Coelho, A. E. (2017). Size, physiological quality, and green seed occurrence influenced by seeding rate in soybeans. *Semina-ciencias Agrarias*, *38*, 595-606.

Fonseca de Oliveira, G. R., Mastrangelo, C. B., Hirai, W. Y., Batista, T. B., Sudki, J. M., Petronilio, A. C. P., Crusciol, C. A. C., & Amaral da Silva, E. A. (2022). An Approach Using Emerging Optical Technologies and Artificial Intelligence Brings New Markers to Evaluate Peanut Seed Quality. *Front Plant Sci*, *13*, 849986. https://doi.org/10.3389/fpls.2022.849986

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14), 2627-2636. https://doi.org/https://doi.org/10.1016/S1352-2310(97)00447-0

Ghasemzadeh, H., Hillman, R. E., & Mehta, D. D. (2024). Toward Generalizable Machine Learning Models in Speech, Language, and Hearing Sciences: Estimating Sample Size and Reducing Overfitting. *Journal of Speech, Language, and Hearing Research*, *67*(3), 753-781. https://doi.org/doi:10.1044/2023_JSLHR-23-00273

Gierz, Ł., Kolankowska, E., Markowski, P., & Koszela, K. (2022). Measurements and Analysis of the Physical Properties of Cereal Seeds Depending on Their Moisture Content to Improve the Accuracy of DEM Simulation. *Applied Sciences*, *12*(2), 549. https://www.mdpi.com/2076-3417/12/2/549

Hankook, H., Lee, S., Kim, K., & Yoo, K. (1990). Comparison Analysis of single Multiplicative neuron with Conventional Neuron Models. *Journal of Theoretical and Applied Information Technology*. https://doi.org/10.1109/ICSPCS.2018.8631758

Kaliniewicz, Z., Markowski, P., Anders, A., Jadwisieńczak, K., Żuk, Z., & Krzysiak, Z. (2019). Physical Properties of Seeds of Eleven Fir Species. *Forests*, *10*(2), 142. https://www.mdpi.com/1999-4907/10/2/142

Kandpal, P., & Mehta, A. (2019). Critical Analysis of Two Dimensional and Four-Dimensional Spiking Neuron Models. *Journal of Computational and Theoretical Nanoscience*. https://doi.org/10.1166/jctn.2019.8268

Kheir, A., Mkuhlani, S., Mugo, J. W., Elnashar, A., Nangia, V., Deware, M., & Govind, A. (2023). Integrating APSIM model with machine learning to predict wheat yield spatial distribution. *Agronomy Journal*. https://doi.org/10.1002/agj2.21470

Mamann, Â. T. W. D., Silva, J. G. d., Binelo, M. O., Scremin, O. B., Kraisig, A. R., Carvalho, I. R., Pereira, L. M., Berlezi, J. D., & Argenta, C. V. (2019). Artificial Intelligence Simulating Grain Productivity During the Wheat Development Considering Biological And Environmental Indicators. *Journal of Agricultural Studies*.

Mittlböck, M., & Heinzl, H. (2002). MEASURES OF EXPLAINED VARIATION IN GAMMA REGRESSION MODELS. *Communications in Statistics - Simulation and Computation*, *31*(1), 61-73. https://doi.org/10.1081/SAC-9687282

Polishchuk, V., & Konovalov, D. V. (2023). The yield of conditioned winter wheat seeds depending on the cultivation technology. *Advanced Agritechnologies*.

Saffariha, M., Jahani, A., & Potter, D. (2020). Seed germination prediction of Salvia limbata under ecological stresses in protected areas: an artificial intelligence modeling approach. *BMC Ecol*, *20*(1), 48. https://doi.org/10.1186/s12898-020-00316-4

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103-113. https://doi.org/https://doi.org/10.1111/j.2041-210X.2010.00012.x

Shamsabadi, E. E. h., Sabouri, H., Soughi, H., & Sajadi, S. J. (2022). Using of Molecular Markers in Prediction of Wheat (Triticum aestivum L.) Hybrid Grain Yield Based on Artificial Intelligence Methods and Multivariate Statistics. *Russian Journal of Genetics, 58*, 603 - 611.

Sieracka, D., Zaborowicz, M., & Frankowski, J. (2023). Identification of Characteristic Parameters in Seed Yielding of Selected Varieties of Industrial Hemp (Cannabis sativa L.) Using Artificial Intelligence Methods. *Agriculture*, *13*(5), 1097. https://www.mdpi.com/2077-0472/13/5/1097

Thangjam, U., & Sahoo, U. K. (2016). Effect of Seed Mass on Germination and Seedling Vigour of Parkia Timoriana (DC.) Merr. *Current Agriculture Research Journal*, *4*, 171-178.

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234-240. https://doi.org/https://doi.org/10.1016/j.sbspro.2013.12.027

Zhang, H., Ji, J., Ma, H., Guo, H., Liu, N., & Cui, H. (2023). Wheat Seed Phenotype Detection Device and Its Application. *Agriculture*, *13*(3), 706. https://www.mdpi.com/2077-0472/13/3/706