



Applying convolutional neural networks for mustard variety recognition

Laura Slebioda · Bogna Zawieja

Received: 16 July 2024 / Accepted: 5 January 2025 / Published online: 16 January 2025
© The Author(s) 2025

Abstract The aim of this study was to develop and apply a Convolutional Neural Network (CNN) model to recognize and classify white mustard (*Sinapis Alba L.*) varieties, addressing the complex task of discriminating among 57 varieties. Utilizing a one-dimensional CNN model, the research focused on multivariate analysis based on a set of 15 traits. The CNN architecture included convolutional layers, batch normalization, pooling, flattening, dropout, and dense layers. The model demonstrated effectiveness in classifying varieties, achieving high accuracy and providing valuable insights into potential new varieties. Subset division, a new approach, was applied. Evaluation metrics, including accuracy, F1 score, precision, and recall, were calculated for eight subsets, confirming the model's robust performance. While this study uses mustard as an illustrative example, the method is not limited to this crop and can be extended to other agricultural crops, with potential modifications depending on the specific traits relevant to each crop. The approach contributes to agricultural advancements, offering a reliable tool for breeders to

assess variety distinctness and streamline the testing process. The model's ability to detect unknown varieties further enhances its utility in agricultural research covering a comprehensive and impactful advancement in variety classification.

Keywords Agricultural algorithms · Classification · Convolutional neural networks · Mustard · Variety recognition

Introduction

The continually growing demand for plant products entails the necessity of breeding new plant varieties that are better adapted to environmental conditions, resistant to diseases, and yielding high crops. Therefore, breeders continually develop new varieties that, after years of research, can be registered. In Poland, breeders submit new varieties to the Research Centre for Cultivar Testing (COBORU). The basis for registration is demonstrating the Distinctness, Uniformity, and Stability (DUS) of the new variety. Variety evaluations are based on two or three years of experiments, where new varieties are sown alongside all known and registered varieties (whose similarity has been indicated by the breeder). Internationally, the coordination of DUS research is managed by the International Union for the Protection of New Varieties of Plant (Union Internationale pur la Protection

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10681-025-03461-3>.

L. Slebioda (✉) · B. Zawieja
Department of Mathematical and Statistical Methods,
Poznań University of Life Sciences, Wojska Polskiego 28,
60-637 Poznań, Poland
e-mail: laura.slebioda@up.poznan.pl

des Obtentions Vegetates – UPOV), headquartered in Geneva.

Distinctness, uniformity, and stability assessments of new varieties traditionally require demonstrating distinctness in at least one attribute. As mentioned above, the new variety must be distinct from known, already registered varieties and other new varieties sown in the same experiment. One-dimensional analysis of variance is conducted separately for each attribute. These principles are recommended by UPOV. However, before conducting separate analyses, it would be worthwhile to apply multivariate methods to determine whether a given variety differs from others simultaneously in all examined attributes. Multivariate evaluations are crucial for selections as they allow for a comprehensive analysis of multiple traits simultaneously, thereby providing a more holistic assessment of a variety's performance (Sharma 2006). If such an analysis shows distinctness of the new varieties, one-dimensional analyses can then be performed to identify specific distinguishing features. If, however, multivariate analysis does not show distinctness of any of the new varieties, there is no need to conduct one-dimensional analyses. Multivariate analysis could be useful for evaluating varieties submitted for registration and by breeders allowing for the evaluation of the novelty of the obtained variety and its chances of a positive outcome in DUS testing. This could eliminate varieties that do not show distinctness from further testing. Yield could also be added to the set of analyzed traits, allowing breeders to identify promising varieties.

In addition to DUS testing, similar methodologies are applicable to the management of germplasm collections, where analogous problems are addressed but with different objectives. Germplasm collections are essential for conserving genetic diversity, which is a critical resource for breeding programs and crop improvement. Accurate classification and differentiation of germplasm accessions are crucial for efficient collection management, as it ensures genetic resources are well-documented and accessible for various research and breeding purposes. For example, Smýkalová et al. (2011) demonstrated how multivariate analysis could be applied in germplasm management to address challenges similar to those faced in DUS testing, such as distinguishing closely related accessions and ensuring effective management strategies. By improving the accuracy and efficiency of

classification, methods like the one proposed in this study can contribute significantly to both DUS testing and germplasm management, advancing agricultural research and breeding efforts.

Mustard is an annual plant belonging to the cabbage family, cultivated for various purposes. These species are often sown as an aftercrop, providing green manure (Alferov and Chernova 2018). Young green plants are valuable fodder for livestock. The seeds are toxic to animals but serve as a culinary spice for humans and for mustard production (Grygier 2023). However, large quantities of seeds are also harmful to humans. White mustard (*Sinapis alba* L.) is used in medicine, such as in the treatment of stomach inflammations, acidity, ulcers, constipation, respiratory diseases (Guo et al. 2013) and many others (Grygier 2023).

In this study, a convolutional neural network (CNN) was used to demonstrate the distinctness of new white mustard varieties. This is a multivariate analysis based on a model that leverages mathematical linear operations between matrices (Albawi, et al., 2017). A CNN model mimics the neurons of the human nervous system. Neural network models take multidimensional data (input data) and through hidden layers, they can output the classification of analyzed objects. CNNs use matrix multiplication in the hidden layers to detect patterns in images or data. The core components of a CNN include convolutional layers, pooling layers, and fully connected layers (Li et al. 2021). Convolutional layers apply filters to the input data to create feature maps, which highlight important features of the data such as edges, textures, and shapes. Pooling layers reduce the dimensionality of these feature maps, retaining the most critical information while reducing computational load and mitigating the risk of overfitting. Fully connected layers, positioned towards the end of the network, take the high-level filtered data and perform the final classification (Zhang et al. 2018). The network's performance depends on the number of its levels. However, as the number of levels increases, the time required for training and testing the network also increases. Therefore, the number of levels should be adjusted to the specific case. 2D-CNN networks, as a biologically inspired model of deep learning, were initially used for image classification, face identification, and pattern recognition (Azizjon, et al. 2020). Convolution operations automatically extract complex feature

patterns from the image using a filter (kernel). In this study, this method was applied to one-dimensional data, meaning 1D-CNN networks were used, consistent with the dimension of white mustard data.

Due to the low values of the evaluation metrics, subset division was applied. In this approach, the original dataset is divided into several subsets. This idea was developed to address challenges associated with distinguishing closely related varieties and improving model sensitivity to specific distinctions. It is a novel concept aimed at helping in situations where the data is limited, making it easier for models to learn from smaller, more focused subsets. While dividing data into subsets is a widely recognized technique in machine learning classification, its application in this study is unique. Abuassba et al. (2017) demonstrated the benefits of subset division by splitting the dataset and applying various classification methods to each subset, ultimately identifying the most effective method for each. However, the approach employed in this study goes beyond their methodology, applying the subset division in a novel context to enhance classification accuracy.

The aim of this work is to develop and apply a CNN model to recognize mustard varieties and to distinguish new mustard varieties from existing ones. Applying a novel subset division approach is to address challenges associated with distinguishing closely related plant varieties, thereby improving the model’s sensitivity to specific distinctions. Although this study focuses on mustard as an example, the developed algorithm is versatile and can be widely applied in agriculture, particularly in optimizing the registration and cultivation of various crops. The approach can be adapted for other crops with appropriate trait data, contributing to broader agricultural advancements.

Materials and methods

Experimental data

The experimental data used in the multivariate analysis are derived from a DUS experiment with white mustard conducted at COBORU in 2009, involving the examination of 57 varieties. The experiments were carried out in a complete randomized block design with two blocks, and for each variety, a sample

of a total of 60 plants was taken. 30 plants were randomly selected from each plot. In this experiment, measurements were taken for 15 quantifiable traits (13 continuous and 2 categorical). The list of analyzed traits is provided in Table 1 below.

It is important to note that decisions about the acceptance or rejection of a variety in DUS tests are typically based on at least 2–3 years of experiments, depending on the species, using the Combined Over Years Uniformity (COYU) method (Kristensen and Roberts 2009). For the single year of data analyzed in this study, the classification of unregistered varieties 8012 and 8040 (that were used to test the ability of the CNN model) was also reviewed based on specific traits identified using the traditional method. Variety 8012 was determined to differ significantly from other varieties based on the traits Plant: Overall length of stem and Siliqua: Length of siliqua. Variety 8040 was found to differ based on the traits Siliqua: Number of seeds in siliqua and Plant: Height at flowering.

Model description

In this study, a Conv1D model, i.e., a one-dimensional convolutional neural network, was employed. Conv1D creates a convolution kernel that is convolved with the input data in one spatial dimension to generate results (Dewantara et al. 2020). Batch Normalization was used in a model for classification

Table 1 List of analyzed traits

Trait Name
Seedling: Length of cotyledons (cm)
Seedling: Width of cotyledons (cm)
Plant: Overall length of stem (cm)
Leaf—Number of leaflets
Leaf: Length of leaf (cm)
Leaf: Width of leaf (cm)
Leaf: Length of leaf petiole (cm)
Plant: Height at flowering (cm)
Flower: Length of flower petal (cm)
Flower: Width of flower petal (cm)
Siliqua: Length of siliqua (excluding beak) (mm)
Siliqua: Length of beak (mm)
Siliqua: Width of siliqua (mm)
Siliqua: Length of pedicel (mm)
Siliqua—Number of seeds in siliqua

purposes. This layer introduces random perturbation to the decision boundary of deep networks, forcing the model to learn boundaries with increased margins to the nearest training samples (Balestrierio and Baraniuk 2022). MaxPooling1D was used to perform pooling operations with the maximum values in a single spatial dimension when creating output data (Dewantara et al. 2020). The flattening layer was utilized as it converts an array into a vector. The Dropout layer was applied as a method to overcome overfitting, which often occurs when using a deep learning approach. The Dropout layer randomly reduces the number of neuron units in the network to decrease the number of connections in each iteration of the learning process. The dense layer that was used, is a simple layer of neurons, where each neuron receives input data from all neurons in the previous layer (Dumane 2020). The model utilized in this study is a sequential model.

The sequential model applied to the analyzed data for mustard varieties is presented in Table 2. It was initialized using the Keras library (an open-source Python library for machine learning). The layers of the model include two convolutional layers (Conv1D), two Batch Normalization layers, a pooling layer (MaxPooling1D), a flattening layer (Flatten), a Dropout layer, and two Dense layers. Each layer has its own parameters such as size, activation function, and input shape. In total, the model has 2,408 parameters, of which 2,344 are trainable, and

64 are not, as they are parameters of batch normalization layers. The activation function in layers 1, 3, and 7 is a rectified linear unit (ReLU). If the input value is below zero, the output value is zero. When the input value rises above a certain threshold, it has a linear relationship with the dependent variable (Bisen 2021). The formula for the ReLU function is provided below (Formula 1). The activation function in layer 9 is Softmax. This function is a combination of multiple sigmoid functions (Sharma, et al. 2020). Since we know that the sigmoid function returns values in the range from 0 to 1, they can be treated as probabilities of data points for a specific class. Unlike sigmoid functions used for binary classification, the Softmax function can be used for multi-class classification problems. The function returns the probability for each data point for all individual classes. When building a network or model for multi-class classification, the output layer of the network will have the same number of neurons as the number of classes in the target object. The formula for the Softmax function is provided below (Formula 2).

Formula 1. ReLU activation function [Bisen 2021].

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

where x is the input to the ReLU activation function.

Formula 2. Softmax activation function [Sharma, et al. 2020].

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \text{ for } j = 1, \dots, K.$$

where:

$\sigma(z)_j$ is the probability for every data point of all the individual classes,

$(z)_j$ is the input to the Softmax activation function corresponding to the j -th element.

In the model, applied loss function was 'categorical_crossentropy,' the optimizer was 'adam,' and the metric was 'accuracy.' The training parameters of the model were configured as follows: the 'epochs' parameter, determining the number of training epochs, is set to 80; the 'batch_size' parameter, defining the size of the training data batches, is set to 50; and applied 'validation_data' technique uses test data as the validation set during training.

Table 2 Model summary (Model: "sequential")

Layer Number	Layer	Output Shape	Parameter
1	conv1d	(None, 13, 16)	64
2	batch_normalization	(None, 13, 16)	64
3	conv1d	(None, 11, 16)	784
4	batch_normalization	(None, 11, 16)	64
5	max_pooling1d	(None, 5, 16)	0
6	flatten	(None, 80)	0
7	dense	(None, 16)	1296
8	dropout	(None, 16)	0
9	dense_1	(None, 8)	136
	Total parameters:	2,408	
	Trainable parameters:	2,344	
	Non-trainable params:	64	

The results for the applied algorithm were not satisfactory, so cross-validation was applied but the results did not improve. It was then decided to use a new approach – division into subsets. It turned out that the best solution is to divide it into 8 subsets. The data was divided into eight subsets with seven or eight varieties per subset. The aim was to improve the classification accuracy of a complex dataset by

creating smaller, more manageable subsets. The division was performed with the goal of optimizing classification performance. Several different configurations were tested and the optimal subset division was selected based on the highest performance observed in each subset.

The detailed operation of data subdivision for the data used in this study is presented in Fig. 1. First, the

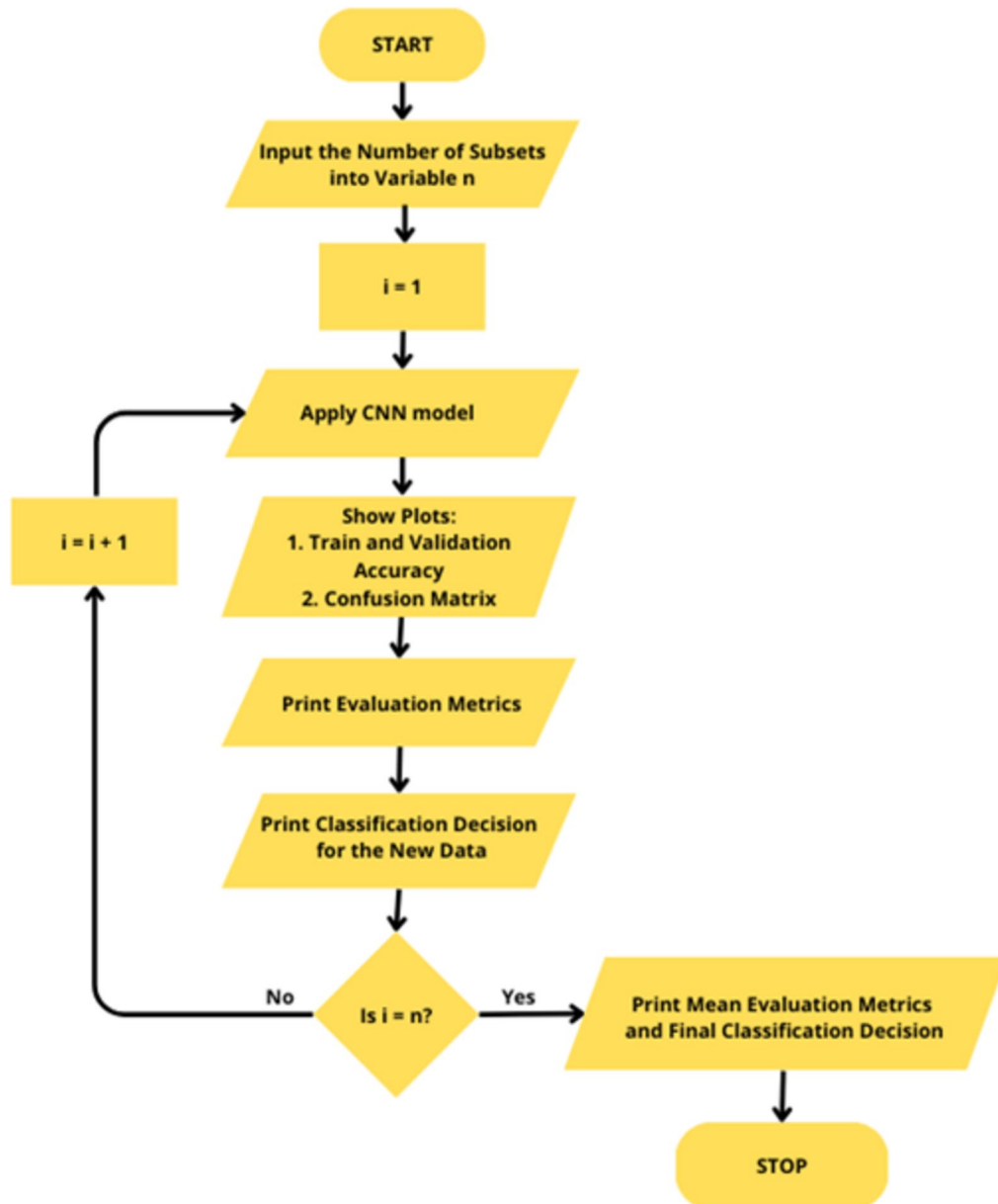


Fig. 1 Flowchart of data subdivision

user inputs the total number of subsets to be processed (variable n) which in our case is eight. Then, the loop counter i is initialized to 1, indicating the start of the loop for processing each subset. For the current subset, a CNN model is applied. This step involves training the model using the subset data. Next, two types of plots are generated and shown. First plot is Train and Validation Accuracy which shows the accuracy of the model on both the training and validation data over epochs. Second plot is Confusion Matrix for the predictions, showing how well the model performed in classifying the data. Then, the evaluation metrics (e.g., accuracy, F1 score, precision, recall) for the current subset are printed. The classification decision for new data is printed, indicating whether the new data is classified as an existing variety or a new variety. A decision point checks if the loop counter i equals n , indicating whether all subsets have been processed. If not, the loop counter i is incremented by 1 and the process loops back to apply the CNN model to the next subset. If i is equal to n , the process proceeds to the final step and the mean evaluation metrics for all subsets are printed and also the final classification decision based on the collective results of all subsets is printed.

It was later explored to incorporate PCA for dimensionality reduction before training the Conv1D model. However, it was observed that while PCA provided a more compact representation of the data, the improvements in model performance were not as substantial as anticipated.

The computations were conducted in Python 3.7.7, utilizing libraries such as tensorflow, scikit-learn, pandas, numpy, scipy, matplotlib, and seaborn.

Application of the model to the data

Subsequently, the created Convolutional Neural Network (CNN) model in the Keras library was applied to each subset. The input data consists of 15 features describing a mustard variety, and the output label is an assigned variety number. Using the StandardScaler method from the sklearn.preprocessing library, the input data was standardized, ensuring that mean is equal to 0 and variance is equal to 1. This standardization process is achieved using the `fit_transform()` method, which calculates mean and standard deviation for each feature based on the available training data and then transforms

the input data by subtracting the mean and dividing by the standard deviation. This process was applied to enhance the stability and effectiveness of machine learning.

Subsequently, using the `'to_categorical'` function from the `tensorflow.keras.utils` library, the output labels are transformed into binary form through a process known as one-hot encoding. This transformation ensures that each label is represented as a binary vector with a length equal to the number of classes, which is required for training a neural network.

Thereafter, the data is divided into a training set and a test set using the `'train_test_split'` function from the `sklearn.model_selection` library. The parameter `'test_size=0.2'` indicates that 20% of the data will constitute the test set. The parameter `'stratify=y'` ensures that the split maintains the class proportions in the output labels. The parameter `'random_state=42'` sets the seed for randomness to ensure reproducibility of the split.

The model is evaluated separately for each subset and in the end, mean is calculated for all evaluation metrics. Additionally, accuracy is calculated on the test set. The evaluation metrics include: train accuracy, test accuracy, F1 score, precision and recall.

The model also assumes the possibility of the existence of previously unknown varieties. When providing the results of assigning the given data to a specific variety, the model may recognize these data as a new variety. Therefore, when submitting data to determine the variety, at least 10 observations should be provided. The model will iterate through all subsets, assign each observation to a specific variety, and then check if each observation is assigned to the same variety. If not, it means that the data is classified as a new variety. After checking this for all subsets, if in any of them, all given observations are classified into one specific variety, then the model classifies the given data as that variety. If in every subset the result is 'The new data is classified as a new variety.', it means that the data belongs to a new, previously unknown variety.

To further validate the proposed method, an additional experiment using a random forest classifier was conducted. The model was built using 100 decision trees with the depth of each tree limited to 10 to avoid overfitting. At each split, only the square root of features was considered to promote diversity among trees and at least 5 samples were required to

split a node, ensuring that each leaf contained at least 2 samples.

Results

Using the trained model, predictions are made on the new data. An experiment was conducted with data related to variety 8012 (10 observations randomly selected from 60 observations). These data were not used previously as training data. The results for each subset are presented in Table 3. Train accuracy ranges from 0.905 to 0.975 and test accuracy ranges from 0.731 to 0.900. The subset where classification for variety 8012 was examined is subset 4. The decision in this subset is: 'The new data belongs to variety: 8012'.

For each subset, charts comparing train accuracy with validation accuracy were made. A representative chart for subset 1 is presented in Fig. 2. The model's accuracy varies depending on the number of epochs and regularization that was conducted using dropout layer. The chart shows that there are no noteworthy differences between the compared accuracies, indicating that overfitting did not occur.

Additionally, a confusion matrix was created for each subset. The overall correct prediction ranges from 58 to 100% (82% on average), with the highest accuracy in subset 4 ranging from 78 to 100% (90% on average). In Fig. 3, a confusion matrix for subset 1 is shown as an illustration.

Mean values for the evaluation metrics are presented in Table 4. Ultimately, the model classified the data as belonging to variety 8012. This indicates that the model is functioning correctly.

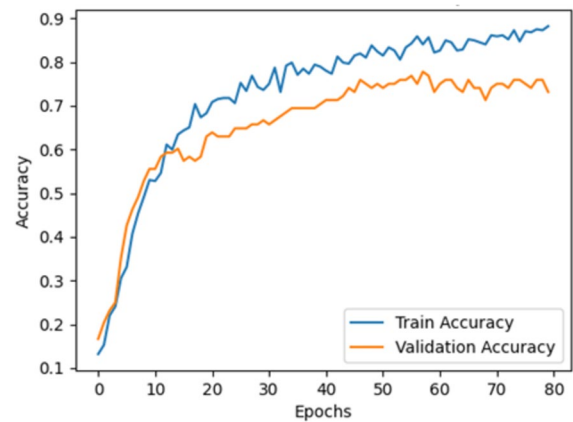


Fig. 2 Chart comparing train accuracy and validation accuracy (subset 1)

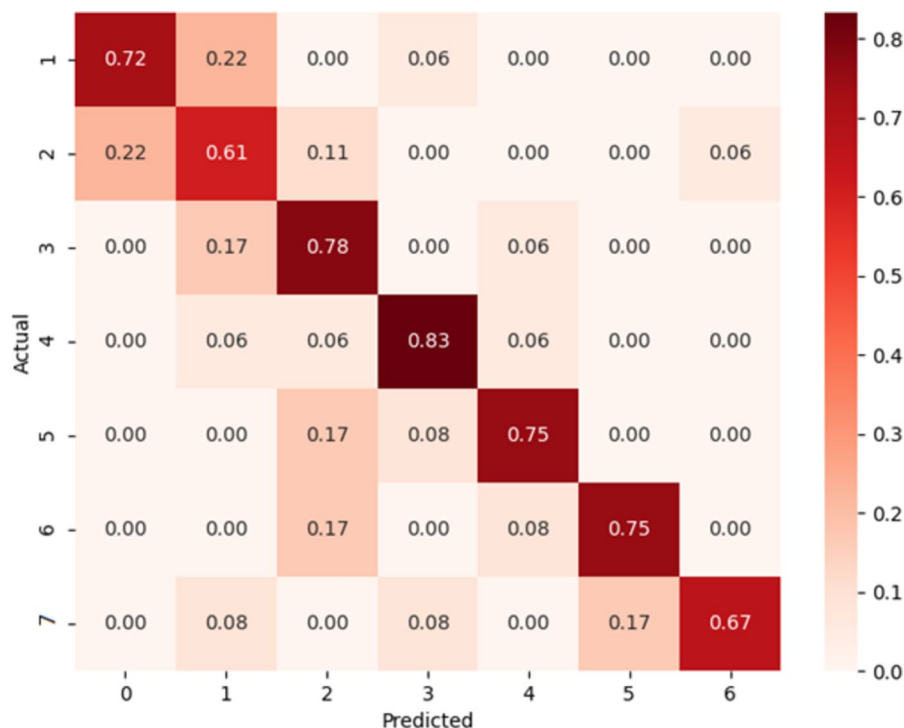
The results of the additional experiment using a random forest classifier are presented in.

Table 5. While random forest model demonstrated slightly better evaluation metrics compared to the 1D CNN, the difference was modest and did not result in a fundamentally different outcome. Notably, both models successfully classified the new data as belonging to variety 8012.

Additionally, model's behavior when encountering entirely unseen varieties was checked. From subset 8, variety 8040 was removed and random 10 observation were used to explore the model's performance in cases where it should classify the input as a "new variety" based on its training experience with the existing subsets. Results are presented in Table 6. 1D CNN model correctly classified the data as a new variety whereas random forest model incorrectly classified the data as belonging to variety 8031.

Table 3 Results for input data of variety 8012

Subset number	Train Accuracy	Test Accuracy	F1 Score	Precision	Recall	Decision
1	0.956	0.731	0.734	0.742	0.731	The new data is classified as a new variety
2	0.975	0.843	0.840	0.845	0.843	The new data is classified as a new variety
3	0.906	0.789	0.793	0.806	0.789	The new data is classified as a new variety
4	0.970	0.900	0.899	0.905	0.900	The new data belongs to variety: 8012
5	0.943	0.774	0.775	0.788	0.774	The new data is classified as a new variety
6	0.905	0.881	0.876	0.877	0.881	The new data is classified as a new variety
7	0.945	0.854	0.851	0.957	0.854	The new data is classified as a new variety
8	0.933	0.824	0.825	0.847	0.824	The new data is classified as a new variety

Fig. 3 Confusion Matrix (subset 1)**Table 4** Mean values for the evaluation metrics and decision

Parameter	Mean
Mean Train Accuracy	0.942
Mean Accuracy	0.824
Mean F1 Score	0.824
Mean Precision	0.833
Mean Recall	0.824
Decision	The new data belongs to variety: 8012

Table 5 Mean values for the evaluation metrics and decision for random forest model

Parameter	Mean
Mean Train Accuracy	0.998
Mean Accuracy	0.875
Mean F1 Score	0.874
Mean Precision	0.879
Mean Recall	0.875
Decision	The new data belongs to variety: 8012

Discussion

While this study specifically addresses the discrimination of plant varieties, many other studies have focused on distinguishing between plant species, which is a less complex task. For example, Lee et al. (2015) conducted a study on convolutional neural networks (CNN) to learn unsupervised feature representations for 44 different plant species. They concluded that learning features through CNN can provide superior feature representation for leaf images compared to hand-crafted features. Saini, et al. (2020) presented a method for classifying healthy and diseased apple and grape leaves using a convolutional neural network. The network achieved a training accuracy of 99.96% and a testing accuracy of 99.90%. Kattenborn, et al. (2019) applied a segmentation approach based on CNN for detailed mapping of plant species communities. The demonstrated approach achieved segmentation and mapping accuracy of at least 84%. Wang, et al. (2021) constructed a multiscale CNN model with attention (AMSCNN) to enhance multiscale feature extraction capabilities. Proposed model was applied to plant species identification and achieved an accuracy of 73.62%. Ashturkar and

Table 6 Results for input data of variety 8040

Subset number	Decision 1D CNN	Decision random forest
1	The new data is classified as a new variety	The new data is classified as a new variety
2	The new data is classified as a new variety	The new data is classified as a new variety
3	The new data is classified as a new variety	The new data is classified as a new variety
4	The new data is classified as a new variety	The new data is classified as a new variety
5	The new data is classified as a new variety	The new data is classified as a new variety
6	The new data is classified as a new variety	The new data is classified as a new variety
7	The new data is classified as a new variety	The new data is classified as a new variety
8	The new data is classified as a new variety	The new data belongs to variety: 8031

Bhalchandra (2021) proposed a designed CNN model and AlexNet model with application-oriented fine-tuning to classify plant species based on leaves. They achieved 91.19% accuracy for their CNN model and 97.41% accuracy for the AlexNet model. Wang and Su (2022) compared CNN models used for detecting the phenotype of four cereal crops. They selected the latest CNNs related to organ counting, weed segmentation, biotic and abiotic stress assessment, and seed variety classification. Results showed the importance of neural network architecture, development strategy, and annotated datasets in designing models for different tasks, directly impacting CNN performance. Despite limited phenotypic information on seeds, the proposed CNN models were able to effectively classify seeds with an accuracy exceeding 95%. The authors also observed that deep CNN models have higher accuracy than shallow networks. Zhu et al. (2018) utilized deep CNN for plant species classification, testing existing classifiers and creating a new one for leaf, flower, fruit, stem, and branch classification. The lowest accuracy was obtained for stem classification, where the new classifier achieved the best result at 65.2% and existing classifiers ranged from 44.8% to 58.7%. The highest accuracy was obtained for fruit classification, with accuracy ranging from 44.8% to 66.3% for known CNN algorithms and 90.2% for the newly created model. Alom et al. (2023) employed CNN for the classification of two types of *B. Napus* species, *B. Rapa* and *B. Alba*, considering flower, leaf, and packet features individually. The DenseNet201 model achieved the best results, getting 100% accuracy for flowers, 97% for leaves, and 100% for packets.

In contrast to previous studies, this article focuses on distinguishing between different varieties. The

presence of a large number of varieties makes classification more challenging and in this study the number of varieties is 57. Morphological characteristics are unstable, they are created as a result of an environmental and climatic conditions, making phenotypic plasticity a consequence of adaptation (Korir et al. 2013). Therefore, molecular markers, statistical tests, and software are employed for classification and characterization of plant varieties. As noted by Korir et al. (2013), certain limitations create challenges to the development and application of modern scientific methods in identifying varieties of many cultivated plants. Taking all of this into consideration, the algorithm developed in this study classifies white mustard varieties very effectively and moreover, informs if a given data belongs to a previously unknown, new variety. Dividing data into subsets turned out to be an efficient method. Distinguishing 57 varieties at once is exceptionally difficult, so the implementation of a new idea, dividing the data into subsets, made this task easier.

While studies on species classification have addressed datasets with a large number of species, the task itself is less complex compared to distinguishing plant varieties due to the finer-grained differences within a single species. In contrast, most studies on variety classification have typically dealt with datasets containing far fewer varieties. For example, Koklu et al. (2021) studied five rice varieties and achieved 100% classification success with CNN, while Wang and Song (2023) examined five varieties of sweet maize seeds using neural networks and achieved 95% accuracy. This study tackles the more complicated task of classifying 57 white mustard varieties, leveraging a novel algorithm to effectively handle such large and complex dataset.

The developed algorithm can be widely used in agriculture, at all stages of cultivation. Thanks to this solution, varieties submitted for registration have the opportunity to achieve positive results in DUS tests and non-promising varieties could be eliminated from further research in terms of distinctness. While random forest model marginally performed better in terms of evaluation metrics, the difference is not significant enough to diminish the utility of the 1D CNN approach. Notably, the 1D CNN demonstrated superior generalization in scenarios involving entirely unseen varieties. When variety 8040 was excluded from the training data and presented as an unseen class, the 1D CNN correctly classified the data as a "new variety." In contrast, the random forest model incorrectly classified the data as belonging to an existing variety (8031). This finding highlights the robustness of the 1D CNN in detecting novel varieties, a critical feature in tasks like variety registration where identifying distinct, previously unobserved classes is essential. It remains a valid approach, as it demonstrated comparable performance while leveraging its convolutional layers to explore potential relationships among adjacent features. This underscores the versatility of both approaches in addressing the task of mustard variety classification.

Surendro et al. (2022) demonstrated that applying PCA to tabular datasets before training Conv1D models can lead to significant improvements in model performance. By reducing the feature space to the most informative components, PCA mitigates the impact of noisy or redundant features, allowing the model to focus on the most relevant patterns. While the experiment conducted in this study did not yield a substantial improvement in performance, this aligns with the understanding that the effectiveness of PCA can vary depending on the specific characteristics of the dataset. Nonetheless, the study by Surendro et al. (2022) supports the notion that PCA, when appropriately applied, can be a powerful preprocessing tool for enhancing deep learning models' effectiveness in tabular data classification tasks. Further investigations into the optimal combination of PCA and Conv1D architecture are needed. In future studies, the proposed approach can be expanded by adding new varieties to the dataset and adjusting the parameters of the CNN architecture for these new varieties. Random forest and other algorithms suitable for tabular data should also be included in future studies

for a comprehensive comparison to provide deeper insights into the advantages and limitations of different approaches.

Author Contribution L.S. did conceptualization, methodology, developed software and prepared all figures. B.Z. was responsible for data curation and supervision. L.S. wrote original draft and B.Z. reviewed and edited it. All authors approved the version to be published.

Funding The publication was financed by the Polish Minister of Science and Higher Education as part of the Strategy of the Poznan University of Life Sciences for 2024-2026 in the field of improving scientific research and development work in priority research areas.

Data availability Data will be available on request.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abuassba AOM, Zhang D, Luo X, Shaheryar A, Ali H (2017) Improving classification performance through an advanced ensemble based heterogeneous extreme learning machines. *Comput Intell Neurosci*. <https://doi.org/10.1155/2017/3405463>
- Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network 2017 International Conference on Engineering and Technology (ICET). <https://doi.org/10.1109/icengtechnol.2017.8308186>

- Alferov AA, Chernova LS (2018) Influence of fertilizers, biomass of white mustard, and biopreparation rhizoagrin on yield and grain quality of spring wheat (*triticum aestivum* L.) and sustainability of the agroecosystem. Russian Agric Sci 44:546–550. <https://doi.org/10.3103/S1068367418060022>
- Alom M, Ali MY, Islam MT, Uddin AH, Rahman W (2023) Species classification of brassica napus based on flowers, leaves, and packets using deep neural networks. J Agric Food Res. <https://doi.org/10.1016/j.jafr.2023.100658>
- Ashturkar HH, Bhalchandra AS (2021) Plant species identification using convolutional neural network intelligent data communication technologies and internet of things. Lect Notes Data Eng Commun Techn. <https://doi.org/10.1007/978-981-15-9509-7>
- Azizjon M, Jumabek A, Kim W (2020) 1D CNN based network intrusion detection with normalization on imbalanced data. 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIC). <https://doi.org/10.1109/icaic48513.2020.9064976>
- Balestrierio R, Baraniuk RG (2022) Batch normalization explained. <https://doi.org/10.48550/arXiv.2209.14778>
- Bisen D (2021) Deep convolutional neural network based plant species recognition through features of leaf. Multimedia Tools Appl 80:6443–6456. <https://doi.org/10.1007/s11042-020-10038-w>
- Dewantara DS, Budi I, Ibrahim MO (2020) 3218IR at SemEval-2020 Task 11: Conv1D and Word Embedding in Propaganda Span Identification at News Articles. Proceedings of the 14th International Workshop on Semantic Evaluation. <https://doi.org/10.18653/v1/2020.semeval-1.225>
- Dumane G (2020) Introduction to Convolutional Neural Network (CNN) using Tensorflow. Towards Data Science. <https://towardsdatascience.com/introduction-to-convolutional-neural-network-cnn-de73f69c5b83> [available: 13.11.2023]
- Grygier A (2023) Mustard seeds as a bioactive component of food. Food Rev Intl 39(7):4088–4101. <https://doi.org/10.1080/87559129.2021.2015774>
- Guo X, Lu H, Lin Y, Chen B, Wu C, Cui Z, Wang Y, Xu Y (2013) Skin penetration of topically applied white mustard extract and its effects on epidermal Langerhans cells and cytokines. Int J Pharm 457(1):136–142. <https://doi.org/10.1016/j.ijpharm.2013.09.015>
- Kattenborn T, Eichel J, Fassnacht FE (2019) Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. Sci Reports. <https://doi.org/10.1038/s41598-019-53797-9>
- Koklu M, Cinar I, Taspinar YS (2021) Classification of rice varieties with deep learning methods. Comput Electron Agric. <https://doi.org/10.1016/j.compag.2021.106285>
- Korir NK, Han J, Shangguan L, Wang C, Kayesh E, Zhang Y, Fang J (2013) Plant variety and cultivar identification: advances and prospects. Crit Rev Biotechnol 33(2):111–125. <https://doi.org/10.3109/07388551.2012.675314>
- Kristensen K, Roberts A (2009) Potential approaches to improving COYU. UPOV Geneva. TWC/27/15: 1–8.
- Lee SH, Chan CS, Wilkin P, Remagnino P (2015) Deep-plant: Plant identification with convolutional neural networks. 2015 IEEE International Conference on Image Processing (ICIP). <https://doi.org/10.1109/icip.2015.7350839>
- Li Z, Liu F, Yang W, Peng S, Zhou J (2021) A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neu Networks Learn Syst. <https://doi.org/10.1109/tnnls.2021.3084827>
- Saini G, Khamparia A, Luhach AK (2020) Classification of Plants Using Convolutional Neural Network. First International Conference on Sustainable Technologies for Computational Intelligence. Advances in Intelligent Systems and Computing. <https://doi.org/10.1007/978-981-15-0029-9>
- Sharma S, Sharma S, Athaiya A (2020) Activation functions in neural networks. Int J Eng Appl Sci Tech 4(12):310–316
- Sharma JR (2006) Statistical and Biometrical Techniques in Plant Breeding. New Age International, New Delhi.
- Smýkalová I, Hampel D, Pavelek M (2011) Determination of duplicates of accessions in a germplasm collection of flax/linseed by means of digital image analysis. Ind Crops Prod 36(1):177–187. <https://doi.org/10.1016/j.indcrop.2011.09.004>
- Surendro K, Rachmatullah M, Santoso J. 2022. Improving 1d Convolutional Neural Network (1d Cnn) Performance in Processing Tabular Datasets Using Principal Component Analysis. Research Square. 8(3): 1
- Wang Y, Song S (2023) Variety identification of sweet maize seeds based on hyperspectral imaging combined with deep learning. Inf Phys Tech. <https://doi.org/10.1016/j.infrared.2023.104611>
- Wang Y-H, Su W-H (2022) Convolutional neural networks in computer vision for grain crop phenotyping: a review. Agronomy. <https://doi.org/10.3390/agronomy12112659>
- Wang X, Zhang C, Zhang S (2021) Multiscale convolutional neural networks with attention for plant species recognition. Comput Intell Neurosci. <https://doi.org/10.1155/2021/5529905>
- Zhang Z, Zhao Y, Liao X, Shi W, Li K, Zou Q, Peng S (2018) Deep learning in omics: a survey and guideline. Brief Funct Genomics. <https://doi.org/10.1093/bfpg/ely030>
- Zhu H, Liu Q, Qi Y, Huang X, Jiang F, Zhang S (2018) Plant identification based on very deep convolutional neural networks. Mult Tools Appl. <https://doi.org/10.1007/s11042-017-5578-9>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.