European Food Research and Technology (2023) 249:3217–3226
https://doi.org/10.1007/s00217-023-04361-3

ORIGINAL PAPER

The classification of leek seeds based on fluorescence spectroscopic data using machine learning

Ewa Ropelewska[1] · **Kadir Sabanci**[2] · **Vanya Slavova**[3] · **Stefka Genova**[4]

Received: 28 June 2023 / Revised: 22 August 2023 / Accepted: 26 August 2023 / Published online: 12 September 2023

Abstract

The objective of this study was to distinguish leek seeds belonging to the *Starozagorski kamush* variety and two breeding lines based on the selected fluorescence spectroscopic data. The classification models were developed for three classes of *Starozagorski kamush* vs. breeding line 4 vs. breeding line 39 and pairs of classes of *Starozagorski kamush* vs. breeding line 4, *Starozagorski kamush* vs. breeding line 39, and breeding line 4 vs. breeding line 39. The traditional machine learning algorithms, such as PART, Logistic, Naive Bayes, Random Forest, IBk, and Filtered Classifier were applied. All three classes were distinguished with an average accuracy of up to 93.33% for models built using IBk and Filtered Classifier. In the case of each model, *Starozagorski kamush* variety was completely different (accuracy of 100%, precision, and F-measure, MCC (Matthews correlation coefficient), and ROC (receiver operating characteristic) area of 1.000) from breeding lines, and the mixing of cases was observed between breeding line 4 and breeding line 39. The models built for pairs of leek seed classes distinguished *Starozagorski kamush* and breeding line 4 with an average accuracy reaching 100% (Logistic, Naive Bayes, Random Forest, IBk). The classification accuracy of *Starozagorski kamush* and breeding line 39 also reached 100% (Logistic, Naive Bayes, Random Forest, IBk), whereas breeding line 4 and breeding line 39 were classified with an average accuracy of up to 80% (Logistic, Naive Bayes, Random Forest, Filtered Classifier). The proposed approach combining fluorescence spectroscopy and machine learning may be used in practice to distinguish leek seed varieties and breeding lines.

Keywords Leek seed variety · Breeding lines · Fluorescence spectroscopy · Classification models · Machine learning algorithms

Introduction

Leek (*Allium porrum* L. or *Allium ampeloprasum* var. *porrum*) is a very important vegetable crop cultivated outdoors all over the world [1–3]. Leek is a biennial herb, related to onion and garlic. It is commonly cultivated as an annual crop [4]. The edible parts of leek are leaves and a bulb. The inflorescence shoot can reach a height of 200 cm. The bicolor of the stem (white shaft with a milder flavor and green shaft with a spicy taste) is related to the presence of different amounts of essential oils [5]. Leek is rich in methyl furan, pentanol, glucosinolates, polysaccharides [6], folic acid, nicotinic acid, lutein, zeaxanthin, vegetable protein, fats, fiber, sulfide oil, minerals, e.g., calcium, zinc, phosphorus, potassium, magnesium, iron, copper, manganese, sodium, and vitamins such as A, B, C, E, and K [5]. As *Allium* species, leek can be a rich source of secondary metabolites, e.g., polyphenolic compounds including flavonoids, phenolic acids, and flavonoid polymers with health benefits. The health-promoting properties of the *Allium* species are associated with organosulfur compounds responsible for the characteristic aroma, taste, and lachrymatory effects [3]. Leek can be characterized by anticancer, antifungal, and antibacterial effects [7]. The consumption of *Allium* vegetables can reduce the risk of colorectal cancer, prostate cancer, breast

✉ Ewa Ropelewska
ewa.ropelewska@inhort.pl

[1] Fruit and Vegetable Storage and Processing Department, The National Institute of Horticultural Research, Konstytucji 3 Maja 1/3, 96-100 Skierniewice, Poland

[2] Department of Electrical and Electronics Engineering, Karamanoglu Mehmetbey University, Karaman, Turkey

[3] Department of Plant Breeding, Agricultural Academy Bulgaria, Maritsa Vegetable Crops Research Institute, 32, Brezovsko Shosse St, 4003 Plovdiv, Bulgaria

[4] Department of Breeding, Maritsa Vegetable Crops Research Institute, Brezovsko Shosse 32, 4000 Plovdiv, Bulgaria

cancer, and or stomach cancer [8]. Due to the presence of various bioactive substances, these vegetables have antioxidant properties [9, 10]. Due to its nutrition and medicinal value, the leek is a culinary and medicinal vegetable [11]. Leek can impart the slight spiciness of a dish and improve its taste. It is a flavor enhancer used in meal preparations and ready-to-heat products. It can be used as a tissue-based system, such as cut leek, and a disrupted system, such as mixed puree-like systems or soups. Leek can be additive to bread, pasta, fermented sausages, and traditional Greek sausages. Dried leek can be added to salads, sauces, soups, meat dishes, and casseroles, and can be an alternative to fresh vegetables [12, 13].

Leek can be characterized by phenotypic and genetic diversity. Different cultivars vary in leaf type and color or stem thickness which can result in different plant morphology and productivity [14]. The growth, yield, and seed characteristics of the leek can be affected by self- and cross-pollination. The properties of leek seeds can also depend on the genotype [1]. *Allium* seeds are black, with rhomboidal or spheroidal shape [15]. Different seed cultivars can be characterized by different properties. Seed quality depends on cultivar purity and distinctness, as well as seed physiological characteristics. Batch purity of high-quality seed cultivars can be essential in marketed species. There are various seed quality control methods useful for cultivar discrimination. Popular sensitive tests, e.g., DNA-based genotyping can be destructive, labor-intensive, complex, and expensive. Therefore, a quality evaluation can be performed for only seeds randomly selected from a batch. Whereas spectroscopy is considered a non-destructive and high-throughput technique for seed evaluation [16], the combination of spectroscopy and chemometric methods may be a promising approach to seed cultivar classification [17]. Furthermore, the procedures combining spectroscopic data with machine learning methods were successfully used for seed quality classification [18].

Fluorescence spectroscopy in the food industry is widely used for quantitative analysis. It is sensitive and specific enough to detect even small concentrations of the compounds [19, 20]. Through it, for example, changes in the structures of proteins, carbohydrates, and lipids in oils can be detected. This is useful for verifying the authenticity of food products [21, 22]. Advances in fiber-optic technology offer outstanding opportunities for the development of a wide range of highly sensitive fiber-optic sensors in many new application areas. Fiber-optic components are successfully adapted to assemblies with micro-optic elements such as lenses, mirrors, prisms, gratings, and others [23, 24]. Fluorescence spectroscopy in agricultural sciences is applied to the analysis of tomatoes [25] and cereals [26]. Their characterization through this technique is performed by grouping objects with similar characteristics to establish methods related to their classification. Until now, the principles of modern optoelectronics have not been used to analyze leek planting material. In the last few years, the demand for high-quality varieties and hybrids of leeks has increased significantly. Therefore, it is important to use non-destructive methods for quality monitoring of leek planting material such as fluorescence spectroscopy [27].

The objective of this study was to distinguish leek seed cultivars using an innovative approach combining fluorescence spectroscopy and machine learning. The application of traditional machine learning algorithms from different groups for the development of models based on selected spectroscopic data to classify leek seed varieties and breeding lines was a great novelty of the present study.

The contributions and prominent features of this manuscript are indicated as follows:

- The application of non-destructive fluorescence spectroscopy for distinguishing leek seed variety and breeding lines.
- Using traditional machine learning algorithms for the classification.
- The development of successful classification models for distinguishing leek seed samples.
- Obtaining high classification accuracies.

Materials and methods

Materials

The samples that are the subject of the study are *Starozagorski kamush*, breeding line number 4 and breeding line number 39. *Starozagorski kamush* is a Bulgarian variety grown throughout the country. It is distinguished by its longer, thin and delicate cylindrical false stem reaching up to 70 cm in height. The leaves are narrow, long light green and upright. Breeding line 4 was created by the inbreeding method in a population of the variety *Starozagorski 72*. It is characterized by a longer false stem of 1.00 m. The leaves are narrow, long, light green and upright. Breeding line 39 was created by breeding offspring of the *Starozagorski* 72 variety with a longer false stem of 90.00 cm. The leaves are narrow, long, dark green and upright.

The seeds were produced at Maritsa Vegetable Crops Research Institute. After removing the leeks in the beginning of November, the cuttings are cleaned by variety, selecting plants typical of the variety. Then they are planted in the field in mid-November with the aim of good rooting. The cuttings are cut at a height of 25 cm and planted in furrows according to the scheme 70/15 for one plant, or 70/30 for three plants in a nest and completely covered with soil. During the growing season, the crop is fed, watered and the phytosanitary status is monitored. Plants develop a single flower stalk. They bloom in July and ripen in September. The seeds are threshed with machines, after which the seeds are cleaned, washed and dried. Drying is carried out in dryers at a temperature of 25–30°C. 20–30 kg/day of seeds are obtained per hectare.

Fluorescence spectroscopy

The study was performed with a fiber-optic spectrometer, which allows the generation of fluorescent emission signals from 200 to 1200 nm. The apparatus is used for performing fluorescence spectroscopy of solid samples at a photosensitive area of 1.9968 × 1.9968 mm. The experimental setup includes a laser diode (emission wavelength 285 nm, optical power 16 mW, DC), portable spectrometer model AvaSpec-ULS2048CL-EVO. The AvaSpec-ULS2048CL-EVO spectrometer is designed for field measurements in the field. The device allows the detection of fluorescent emission signals in any environment regardless of its illumination. By tuning the spectrometer from AvaSoft8, the light signals from the working environment are eliminated and only the useful emission signal of the investigated sample remains. Because of this advantage, the AvaSpec-ULS2048CL-EVO has no requirement for an environment in which to conduct fluorescence measurements with it, and no requirement for an illumination level. The sample is placed on a duralumin stand, which allows the reception of an emission signal from it below 180° by a U-shaped optical fiber. This reduces aberrations and allows the generation of a better quality emission fluorescent signal (Fig. 1). The resolution of the spectrometer can be in the range of 0.06–20 nm, and that of the setting used for our experiment is 0.06 nm. Since the fluorescence is often very weak and also in all directions, in order not to saturate the receiver, the useful fluorescence signal is measured in a direction that is below 180° to the excitation radiation. It is preferable to use a laser diode (LED) as a source in the circuit, as its spectral width is very small. The LED used in the experiment has a relatively wide spectral radiation width of about 30–40 nm and the angular distribution of its radiation is in a large angular range of ± 30°. The sensitivity of the spectrometer is in the range of 200 nm to 1200 nm. Its resolution is $\delta\lambda = 5$ nm. The spectral installation based on fluorescent signals will make it possible to record both the emission spectrum and the spectrum of the excitation source. The emission spectrum is the wavelength distribution of the emission measured for a constant excitation wavelength. The excitation spectrum is the dependence of the emission intensity measured for one wavelength on scanning on the excitation wavelength. This spectrum is represented as a dependence of the wavelength of light on the light intensity incident on the photodetector in the spectrometer.

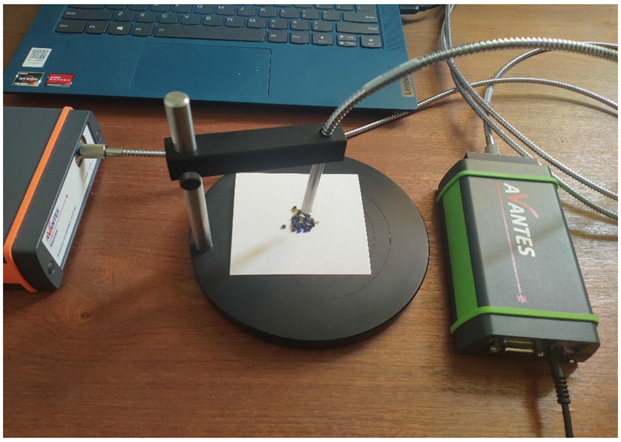

Fig. 1 General view of the experimental installation used by fluorescence spectroscopy

For the specific circuit, the photodetector is of the CMOS type model S9132. Its sensitivity is in the range of 200–1200 nm. Its resolution is $\delta\lambda = 5$ nm. S9132 was chosen because it can detect emission radiation from a sample of garlic with a very low loss of water content due to a false stem grown from a vegetative bud of very short length.

The laser radiation is removed from the source and falls on the sample. The samples represent 5 g of planting material from five different packets containing seeds of *Starozagorski kamush*, breeding line 4 and breeding line 39 located on an area with a radius of 1 cm, which are at a distance of 2 cm from the optical fiber. After the sample fluoresces, the emission signal falls on a U-shaped optical fiber with a core diameter of 200 μm with a step index of refractive index and a numerical aperture of 0.22. The same U-shaped fiber was used to detect all emission signals from the planting material samples and lead the signal to the detector. It takes it to the detector. In the spectrometer, the light signal is converted to electrical-digital via a USB 2.0 wire, downloaded to a computer with AvaSoft8 software and exported to Excel. This allows analysis, processing and visualization of the results of the study.

Five replicates of emission fluorescence signal detection were performed for each seed type: *Starozagorski kamush*, breeding line 4 and breeding line 39. There are five graphs each of *Starozagorski kamush*, breeding line 4 and breeding line 39 (Fig. 2). A difference in the emission fluorescence signal of *Starozagorski kamush* and breeding line 4 as well as breeding line 39 is clearly observed. The spectral wavelength shift and signal intensity level are due to a difference in the content of biologically active substances of a specific variety.

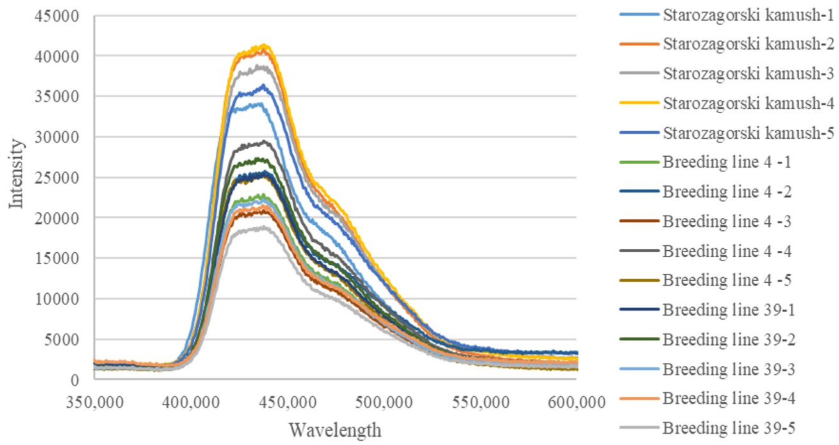

Fig. 2 Difference in emission wavelength for *Starozagorski kamush*, breeding line 4 and breeding line 39

Leek seed classification

The leek seeds belonging to the variety *Starozagorski kamush*, breeding line 4 and breeding line 39 were classified using the WEKA application (Machine Learning Group, University of Waikato, Hamilton, New Zealand) [28–30] based on the fluorescence spectroscopic data. The classification models were built for all three classes and the comparison of pairs, such as *Starozagorski kamush* vs. breeding line 4, *Starozagorski kamush* vs. breeding line 39, and breeding line 4 vs. breeding line 39. The applied procedure is presented in Fig. 3.

For each dataset, attribute selection using the best first with the CFS (correlation-based feature selection) was carried out to choose spectroscopic data, the most useful to distinguish leek seed samples. The models were built based on selected data using a tenfold cross-validation mode and PART (group of Rules), Logistic (group of Functions), Naive Bayes (group of Bayes), Random Forest (group of Trees), IBk (group of Lazy), and Filtered Classifier (group of Meta) traditional machine learning algorithms. The following main parameters of the algorithms were used:

PART-confidenceFactor: 0.25; doNotCheckCapabilities: False; debug: False; batchSize: 100; unpruned: False; minNumObj: 2; numFolds: 3; useMDLcorrection: True; seed: 1,

Logistic-debug: False; doNotCheckCapabilities: False; batchSize: 100; useConjugateGradientDescent: False; ridge: 1.0E-8,

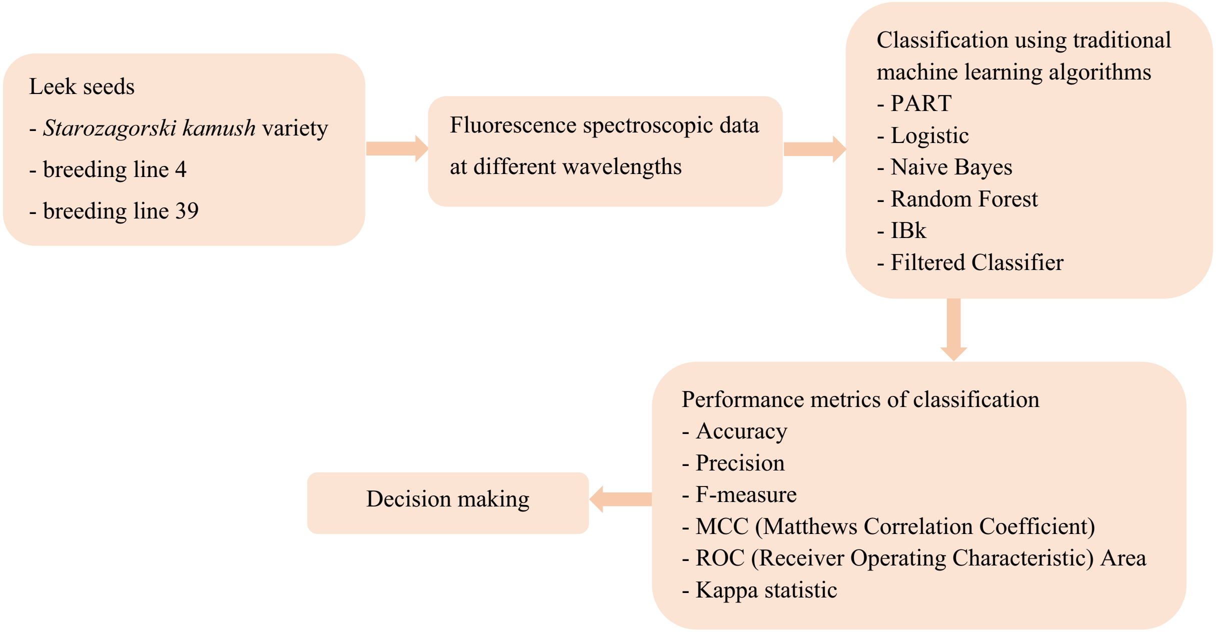
Leek seeds
- Starozagorski kamush variety
- breeding line 4
- breeding line 39
Fluorescence spectroscopic data at different wavelengths
Classification using traditional machine learning algorithms
- PART
- Logistic
- Naive Bayes
- Random Forest
- IBk
- Filtered Classifier
Performance metrics of classification
- Accuracy
- Precision
- F-measure
- MCC (Matthews Correlation Coefficient)
- ROC (Receiver Operating Characteristic) Area
- Kappa statistic
Decision making

Fig. 3 The procedure applied to classify leek seed variety and breeding lines based on fluorescence spectroscopic data using machine learning algorithms

Naive Bayes-debug: False; doNotCheckCapabilities: False; batchSize: 100; displayModelInOldFormat: False; useSupervisedDiscretization: False; useKernelEstimator: False,

Random Forest-doNotCheckCapabilities: False; batchSize: 100; breakTiesRandomly: False; debug: False; numIterations: 100; numExecutionSlots: 1; seed: 1,

IBk-KNN: 1; doNotCheckCapabilities: False; batchSize: 100; nearestNeighbourSearchAlgorithm: LinearNNSearch–distanceFunction: Euclidean Distance-R first-last; debug: False; meanSquared: False; windowSize: 0,

Filtered Classifier-batchSize: 100; classifier: J48-C 0.25-M 2; doNotCheckCapabilities: False; debug: False; filter: Discretize-R firs-last—precision 6; seed: 1.

The number of correctly and incorrectly classified cases, average accuracy, and the values of precision, F-measure, MCC (Matthews correlation coefficient), ROC (receiver operating characteristic) area, and Kappa statistic were determined [31–33] (Eqs. 1–8).

$$Accuracy = \frac{(TP + TN)}{TP + TN + FN + FP}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = TPR = \frac{TP}{TP + FN}, \tag{3}$$

$$FPR = \frac{FP}{FP + TN}, \tag{4}$$

$$F - measure = \frac{2 * precsion * recall}{(precision + recall)}, \tag{5}$$

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}}, \tag{6}$$

$$ROCarea = areaunderTPRvs.FPRcurve, \tag{7}$$

$$Kappa = \frac{\frac{(TP+FP)(TP+FN)}{(TP+FP)(TP+FN)(TN+FP)(TN+FN)} + \frac{(TN+FP)(TN+FN)}{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}, \tag{8}$$

where TP: true positive; TN: true negative; FP: false positive; FN: false negative.

Results

This section discusses machine learning-based analysis of leek seed spectroscopic data. In this context, leek seeds of two breeding lines (4 and 36) and *Starozagorski kamush* variety were distinguished by using six different machine learning algorithms. To examine the performance of several approaches for sorting leek seeds, the tables below include the confusion matrix, precision, Kappa statistic, ROC area, F-measure, MCC, and average accuracy metrics.

For distinguishing all three samples (two breeding lines and one variety), the performances of PART, Logistic, Naive Bayes, Random Forest, IBk, and Filtered Classifier machine learning algorithms are shown in Table 1. The average accuracy reached 93.33% in the case of a classification model built using the IBk and Filtered Classifier algorithms. The Kappa statistic was equal to 0.90. 10% of cases from breeding line 4 were classified as breeding line 36, and 10% of cases belonging to breeding line 36 were classified as breeding line 4. The model developed using IBk correctly classified all cases of *Starozagorski kamush*. For the Filtered Classifier, all leek seeds belonging to *Starozagorski kamush* and breeding line 39 were correctly classified, whereas 20% of cases from the actual class of breeding line 4 were incorrectly included in the predicted class breeding line 36. The great mixing of cases between classes was observed for the PART algorithm, which achieved an average accuracy of 86.67% for distinguishing three different samples of leek seeds. When the confusion matrices for the classification conducted using PART were evaluated, it was noted that all of the *Starozagorski kamush* samples were accurately distinguished. However, the PART classifier incorrectly classifies 10% of leek seeds from breeding line 4 as *Starozagorski kamush* and 20% of breeding line 4 as breeding line 39. In addition, the PART classifier incorrectly classifies 10% of leek seeds from breeding line 39 as breeding line 4. For the other machine learning algorithms, the *Starozagorski kamush* seeds were correctly distinguished from other classes and the mixing of cases occurred between breeding lines. The values of precision, F-measure, MCC, and ROC area were the highest for *Starozagorski kamush* and reached 1.000 in the case of Logistic, Naive Bayes, Random Forest, IBk, and Filtered Classifier.

The results of distinguishing *Starozagorski kamush* and breeding line 4 are presented in Table 2. The completely correct classification with accuracies of 100%, Kappa statistic of 1.00 and precision, F-measure, MCC, and ROC area of 1.000 was obtained for models built using Logistic, Naive Bayes, Random Forest, and IBk. The average accuracy equal to 85% was the lowest for a model built using PART. 10% of cases belonging to *Starozagorski kamush* were incorrectly classified as breeding line 4 and 20% from breeding line 4 were incorrectly classified as *Starozagorski kamush*.

Slightly higher accuracies were obtained for the classification of leek seeds *Starozagorski kamush* and breeding line 39 (Table 3). For the models developed using Logistic, Naive Bayes, Random Forest, and IBk, both classes were completely correctly distinguished in 100%. Whereas the

Table 1 The performance metrics of classification of leek seeds belonging to *Starozagorski kamush*, breeding line 4 and breeding line 39 based on fluorescence spectroscopic data

Algorithm	Predicted class (%)			Actual class	Average accuracy (%)	Precision	F-measure	MCC	ROC area	Kappa statistic
	Starozagorski kamush	Breeding line 4	Breeding line 39							
Rules. PART	100	0	0	*Starozagorski kamush*	86.67	0.909	0.952	0.929	0.975	0.80
	10	70	20	Breeding line 4		0.875	0.778	0.693	0.813	
	0	10	90	Breeding line 39		0.818	0.857	0.783	0.883	
Functions. Logistic	100	0	0	*Starozagorski kamush*	86.67	1.000	1.000	1.000	1.000	0.80
	0	70	30	Breeding line 4		0.875	0.778	0.693	0.850	
	0	10	90	Breeding line 39		0.750	0.818	0.722	0.893	
Bayes. Naive Bayes	100	0	0	*Starozagorski kamush*	90	1.000	1.000	1.000	1.000	0.85
	0	80	20	Breeding line 4		0.889	0.842	0.772	0.915	
	0	10	90	Breeding line 39		0.818	0.857	0.783	0.935	
Trees. Random Forest	100	0	0	*Starozagorski kamush*	90	1.000	1.000	1.000	1.000	0.85
	0	80	20	Breeding line 4		0.889	0.842	0.772	0.905	
	0	10	90	Breeding line 39		0.818	0.857	0.783	0.920	
Lazy. IBk	100	0	0	*Starozagorski kamush*	93.33	1.000	1.000	1.000	1.000	0.90
	0	90	10	Breeding line 4		0.900	0.900	0.850	0.925	
	0	10	90	Breeding line 39		0.900	0.900	0.850	0.925	
Meta. Filtered Classifier	100	0	0	*Starozagorski kamush*	93.33	1.000	1.000	1.000	1.000	0.90
	0	80	20	Breeding line 4		1.000	0.889	0.853	0.905	
	0	0	100	Breeding line 39		0.833	0.909	0.866	0.930	

MCC Matthews correlation coefficient, *ROC area* receiver operating characteristic area

Table 2 The classification of leek seeds *Starozagorski kamush* and breeding line 4

Algorithm	Predicted class (%)		Actual class	Average accuracy (%)	Precision	F-measure	MCC	ROC area	Kappa statistic
	Starozagorski kamush	Breeding line 4							
Rules. PART	90	10	*Starozagorski kamush*	85	0.818	0.857	0.704	0.850	0.70
	20	80	Breeding line 4		0.889	0.842	0.704	0.850	
Functions. Logistic	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 4		1.000	1.000	1.000	1.000	
Bayes. Naive Bayes	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 4		1.000	1.000	1.000	1.000	
Trees. Random Forest	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 4		1.000	1.000	1.000	1.000	
Lazy. IBk	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 4		1.000	1.000	1.000	1.000	
Meta. Filtered Classifier	90	10	*Starozagorski kamush*	90	0.900	0.900	0.800	0.900	0.80
	10	90	Breeding line 4		0.900	0.900	0.800	0.900	

MCC Matthews correlation coefficient, *ROC area* receiver operating characteristic area

lowest average accuracy of 90% was observed for a model built using Filtered Classifier. In the case of this model, both classes were correctly classified in 90%, and 10% belonging to each class were incorrectly classified as the second class. The Kappa statistic was equal to 0.80. The values of precision of 0.900, F-measure of 0.900, MCC of 0.800, and ROC area of 0.900 were determined for both classes.

The least correct classification and the greatest mixing of cases were found for the distinguishing breeding line 4 and breeding line 39 (Table 4). An average accuracy reaching 80% and the Kappa statistic equal to 0.60 were found

Table 3 The distinguishing leek seeds *Starozagorski kamush* and breeding line 39

Algorithm	Predicted class (%)		Actual class	Average accuracy (%)	Precision	F-measure	MCC	ROC area	Kappa statistic
	Starozagorski kamush	Breeding line 39							
Rules. PART	100	0	*Starozagorski kamush*	95	0.909	0.952	0.905	0.950	0.90
	10	90	Breeding line 39		1.000	0.947	0.905	0.950	
Functions. Logistic	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 39		1.000	1.000	1.000	1.000	
Bayes. Naive Bayes	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 39		1.000	1.000	1.000	1.000	
Trees. Random Forest	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 39		1.000	1.000	1.000	1.000	
Lazy. IBk	100	0	*Starozagorski kamush*	100	1.000	1.000	1.000	1.000	1.00
	0	100	Breeding line 39		1.000	1.000	1.000	1.000	
Meta. Filtered Classifier	90	10	*Starozagorski kamush*	90	0.900	0.900	0.800	0.900	0.80
	10	90	Breeding line 39		0.900	0.900	0.800	0.900	

MCC Matthews correlation coefficient, *ROC area* receiver operating characteristic area

Table 4 The classification of leek seeds belonging to breeding line 4 and breeding line 39

Algorithm	Predicted class (%)		Actual class	Average accuracy (%)	Precision	F-measure	MCC	ROC area	Kappa statistic
	Breeding line 4	Breeding line 39							
Rules. PART	70	30	Breeding line 4	75	0.778	0.737	0.503	0.775	0.50
	20	80	Breeding line 39		0.727	0.762	0.503	0.775	
Functions. Logistic	80	20	Breeding line 4	80	0.800	0.800	0.600	0.820	0.60
	20	80	Breeding line 39		0.800	0.800	0.600	0.860	
Bayes. Naive Bayes	90	10	Breeding line 4	80	0.750	0.818	0.612	0.880	0.60
	30	70	Breeding line 39		0.875	0.778	0.612	0.880	
Trees. Random Forest	90	10	Breeding line 4	80	0.750	0.818	0.612	0.880	0.60
	30	70	Breeding line 39		0.875	0.778	0.612	0.880	
Lazy. IBk	70	30	Breeding line 4	70	0.700	0.700	0.400	0.700	0.40
	30	70	Breeding line 39		0.700	0.700	0.400	0.700	
Meta. Filtered Classifier	80	20	Breeding line 4	80	0.800	0.800	0.600	0.830	0.60
	20	80	Breeding line 39		0.800	0.800	0.600	0.830	

MCC Matthews correlation coefficient, *ROC area* receiver operating characteristic area

for models developed by Logistic, Naive Bayes, Random Forest, and Filtered Classifier. The highest accuracy for individual classes was equal to 90% for breeding line 4 in the case of Naive Bayes and Random Forest, whereas leek seeds belonging to breeding line 39 were correctly classified in 70% for these algorithms. The remaining cases were incorrectly included in other classes. The model built using IBk was characterized by the lowest average accuracy of 70%. Both breeding lines were correctly classified in 70%. The Kappa statistic was 0.40 and the precision of 0.700, F-measure of 0.700, MCC of 0.800, and ROC area of 0.700 for both classes were observed.

The performed study revealed the usefulness of fluorescence spectroscopic data for distinguishing leek seed varieties and breeding lines using machine learning models. Spectroscopic techniques are effective in seed research and can be used in various aspects. For example, da Silva Medeiros et al. [34] applied near-infrared spectroscopy (NIRS) to discriminate *Brassica* seed species with correctness of 94.9%. The usefulness of near-infrared (NIR) spectroscopy for sexing papaya seeds with an F-score value equal to 0.81 was also confirmed [35]. Furthermore, NIR spectroscopy proved to be an effective technique for the assessment of insect infestation and protein content of maize seeds [36] and the quantification of phenolic content and total flavonoids in raw peanut seeds [37]. Terahertz spectroscopy was used for the recognition of transgenic cotton seeds [38] and the identification of the adulteration of rice seeds [39]. In view of the promising literature data, various spectroscopic techniques can be used in future studies in various directions of leek seed quality assessment.

Conclusions

The performed study involved a novel approach combining fluorescence spectroscopy and traditional machine learning algorithms to distinguish leek seed varieties and breeding lines. The applied procedure was innovative on the background of available literature for the assessment of leek seed quality. Machine learning models built based on spectroscopic data allowed for the classification of three seed types with an accuracy reaching 93.33%. The most effective algorithms were IBk and Filtered Classifier. Each model distinguished *Starozagorski kamush* variety with breeding lines 39 and 4 with the highest accuracy, whereas the greatest mixing of leek seeds was found between breeding lines. Future studies can involve also other spectroscopic techniques in leek seed studies or various directions of seed quality assessment including the discrimination of species, varieties, and breeding lines, as well as the estimation of the content of chemical compounds in seeds.

Data availability The data presented in this study are available upon request from the corresponding author.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

Compliance with ethics requirements This article does not contain any studies with human or animal subjects.

References

1. De Clercq H, Peusens D, Roldan-Ruiz I, Van Bockstaele E (2003) Causal relationships between inbreeding, seed characteristics and plant performance in leek (*Allium porrum* L.). Euphytica 134:103–115
2. Appeltans S, Pieters JG, Mouazen AM (2021) Detection of leek rust disease under field conditions using hyperspectral proximal sensing and machine learning. Remote Sens 13:1341
3. Bernaert N, Debonne E, De Leyn I, Van Droogenbroeck B, Van Bockstaele F (2022) Incorporation of leek powder (*Allium ampeloprasum* var. *porrum*) in wheat bread: technological implications, shelf life and sensory evaluation. LWT-Food Sci Technol 153:112517
4. Fattorusso E, Lanzotti V, Taglialatela-Scafati O, Di Rosa M, Ianaro A (2000) Cytotoxic saponins from bulbs of *Allium porrum* L. J Agric Food Chem 48:3455–4346
5. Biernacka B, Dziki D, Kozłowska J, Kowalska I, Soluch A (2021) Dehydrated at different conditions and powdered leek as a concentrate of biologically active substances: antioxidant activity and phenolic compound profile. Materials 14:6127
6. Sanchez-Salvador JL, Marques MP, Brito MSCA, Negro C, Monte MC, Manrique YA, Santos RJ, Blanco A (2022) Valorization of vegetable waste from leek, lettuce, and artichoke to produce highly concentrated lignocellulose micro- and nanofibril suspensions. Nanomaterials 12:4499
7. Baky MH, Shamma SN, Khalifa MR, Farag MA (2023) How does allium leafy parts metabolome differ in context to edible or inedible taxa? case study in seven allium species as analyzed using ms-based metabolomics. Metabolites 13:18
8. Kratchanova M, Nikolova M, Pavlova E, Yanakieva I, Kussovski V (2010) Composition and properties of biologically active pectic polysaccharides from leek (*Allium porrum*). J Sci Food Agric 90:2046–2051
9. Golisz E, Wielewska I, Roman K, Kacprzak M (2022) Probabilistic model of drying process of leek. Appl Sci 12:11761
10. Sałata A, Nurzyńska-Wierdak R, Kalisz A, Moreno-Ramón H (2022) Impacts of alexandrian clover living mulch on the yield, phenolic content, and antioxidant capacity of leek and shallot. Agronomy 12:2602
11. Wang Y, Li X, Shen J, Lang H, Dong S, Zhang L, Fang H, Yu Y (2022) Uptake, translocation, and metabolism of thiamethoxam in soil by leek plants. Environ Res 211:113084
12. Biernacka B, Dziki D, Gawlik-Dziki U (2022) Pasta enriched with dried and powdered leek: physicochemical properties and changes during cooking. Molecules 27:4495
13. Delbaere SM, Bernaerts T, Vancoillie F, Buvé C, Hendrickx ME, Grauwet T, Van Loey AM (2022) Comparing the effect of several pretreatment steps, selected to steer (bio) chemical reactions, on the volatile profile of leek (*Allium ampeloprasum* var. *porrum*). LWT-Food Sci Technol 172:114205
14. Melouk SAM, Hassan MA, Elwan MWM, El-Seifi SK, Habib ES, Yousef EAA (2023) Horticultural, chemical and genetic diversity using SSR markers in Leek germplasm collection. Sci Hortic 311:111782
15. Vuković S, Popović-Djordjević JB, Kostić AŽ, Pantelić ND, Srećković N, Akram M, Laila U, Katanić Stanković JS (2023) *Allium* species in the balkan region—major metabolites. Antioxid Antimicrob Prop Hortic 9:408
16. Reddy P, Panozzo J, Guthridge KM, Spangenberg GC, Rochfort SJ (2023) Single seed near-infrared hyperspectral imaging for classification of perennial ryegrass seed. Sensors 23:1820
17. Shrestha S, Deleuran LCh, Gislum R (2016) Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics. J Spectr Imaging 5:1–8
18. Medeiros ADd, Silva LJd, Ribeiro JPO, Ferreira KC, Rosas JTF, Santos AA, Silva CBd (2020) Machine learning for seed quality classification: an advanced approach using merger data from ft-nir spectroscopy and x-ray imaging. Sensors 20:4319
19. Qin J, Lu R (2008) Measurement of the optical properties of fruits and vegetables using spatially resolved hyperspectral diffuse reflectance imaging technique. Postharvest Biol Technol 49:355–365
20. Valeur B, Santos B, Molecular M (2012) Fluorescence: principles and applications. Wiley-VCH, pp 13–19
21. Bachmann L, Zezell DM, Ribeiro AdC, Gomes L, Ito AS (2006) Fluorescence spectroscopy of biological tissues. A Rev Appl Spectrosc Rev 41:575–590
22. Hof M, Hutterer R, Fidler V (2005) Fluorescence spectroscopy in biology. Springer, Cham, pp 91–182
23. Dakin J, Brown R (2006) Handbook of Optoelectronics. CRC Press, pp 74–253
24. Mitchke F (2010) Fiber optics physics and technology Heidelberg. Springer, pp 47–103
25. Hoffmann A, Noga G, Hunsche M (2015) Fluorescence indices for monitoring the ripening of tomatoes in pre- and postharvest phases. Sci Hortic 191:74–81
26. Blecker C (2011) Fluorescence spectroscopy measurement for quality assessment of food systems—a review. Food Bioprocess Technol 4:364–386
27. Hyde P, Mutschler EE, M, (2012) Doubled haploid onion (*Allium cepa* L.) lines and their impact on hybrid performance. HortScience 47:1690–1695
28. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, Scuse D (2016) WEKA manual for version 3-9-1. University of Waikato, Hamilton, New Zealand
29. Frank E, Hall M, Witten I (2016) Online appendix for "data mining: practical machine learning tools and techniques", the WEKA workbench. Elsevier, Amsterdam, The Netherlands
30. Witten IH, Frank E, Hal, MA, Pal CJ 2005 Practical machine learning tools and techniques: In proceedings of the data mining, Las Vegas, NV, USA, pp 20–23
31. Ropelewska E, Szwejda-Grzybowska J (2021) A comparative analysis of the discrimination of pepper (Capsicum annuum L.) based on the cross-section and seed textures determined using image processing. J Food Process Eng 44:e13694
32. Sabanci K, Aslan MF, Ropelewska E, Unlersen MF (2021) A convolutional neural network-based comparative study for pepper seed classification: analysis of selected deep features with support vector machine. J Food Process Eng 45:e13955
33. Ropelewska E, Rady AM, Watson NJ (2023) Apricot stone classification using image analysis and machine learning. Sustainability 15:9259
34. da Silva Medeiros ML, Cruz-Tirado JP, Lima AF, de Souza Netto JM, Ribeiro APB, Bassegio D, Godoy HT, Barbin DF (2022) Assessment oil composition and species discrimination of Brassicas seeds based on hyperspectral imaging and portable near

infrared (NIR) spectroscopy tools and chemometrics. J Food Compos Anal 107:104403

35. Silva Fernandes TF, de Oliveira Silva RV, de Freitas DLD, Sanches AG, da Silva M, Cunha Júnior LC, de Lima KG, de Almeida Teixeira GH (2022) Sex type determination in papaya seeds and leaves using near infrared spectroscopy combined with multivariate techniques and machine learning. Comput Electron Agric 193:106674
36. Wang Z, Huang W, Li J, Liu S, Fan S (2023) Assessment of protein content and insect infestation of maize seeds based on online near-infrared spectroscopy and machine learning. Comput Electron Agric 211:107969
37. Haruna SA, Li H, Wei W, Geng W, Luo X, Zareef M, Yao-Say Solomon Adade S, Ivane NMA, Isa A, Chen Q (2023) Simultaneous quantification of total flavonoids and phenolic content in raw peanut seeds via NIR spectroscopy coupled with integrated algorithms. Spectrochim Acta Part A Mol Biomol Spectrosc 285:121854
38. Yi C, Tuo S, Zhang L, Xiao H (2022) Improved kernel entropy composition analysis method for transgenic cotton seeds recognition based on terahertz spectroscopy. Chemom Intell Lab Syst 225:104575
39. Hou X, Jie Z, Wang J, Liu X, Ye N (2023) Application of terahertz spectroscopy combined with feature improvement algorithm for the identification of adulterated rice seeds. Infrared Phys Technol 131:104694

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