





OPEN

## Ensemble and optimization algorithm in support vector machines for classification of wheat genotypes

Mujahid Khan<sup>1,2</sup>, B. K. Hooda<sup>2</sup>, Arpit Gaur<sup>3,5</sup>, Vikram Singh<sup>3</sup>, Yogesh Jindal<sup>3</sup>, Hemender Tanwar<sup>4</sup>, Sushma Sharma<sup>4</sup>, Sonia Sheoran<sup>5</sup>, Dinesh Kumar Vishwakarma<sup>6</sup> , Mohammad Khalid<sup>7</sup>, Ghadah Shukri Albakri<sup>8</sup>, Maha Awjan Alreshidi<sup>9</sup>, Jeong Ryeol Choi<sup>10</sup>  & Krishna Kumar Yadav<sup>11,12</sup>

This study aimed to classifying wheat genotypes using support vector machines (SVMs) improved with ensemble algorithms and optimization techniques. Utilizing data from 302 wheat genotypes and 14 morphological attributes to evaluate six SVM kernels: linear, radial basis function (RBF), sigmoid, and polynomial degrees 1–3. Various optimization methods, including grid search, random search, genetic algorithms, differential evolution, and particle swarm optimization, were used. The radial basis function kernel achieves the highest accuracy at 93.2%, and the weighted accuracy ensemble further improves it to 94.9%. This study shows the effectiveness of these methods in agricultural research and crop improvement. Notably, optimization-based SVM classification, particularly with particle swarm optimization, saw a significant 1.7% accuracy gain in the test set, reaching 94.9% accuracy. These findings underscore the efficacy of RBF kernels and optimization techniques in improving wheat genotype classification accuracy and highlight the potential of SVMs in agricultural research and crop improvement endeavors.

**Keywords** Ensemble algorithm, Ensemble weighted average (EWA), Wheat genotypes classification, Radial basis function, Support vector machine

Machine learning (ML) is a multi-disciplinary stream which builds on concepts from various other branches like computer science, cognitive science, optimization, statistics and mathematics<sup>1</sup>. The analysis and interpretation of data for accurate prediction and classification has been an important field of research from several decades in machine learning. Classification problems can be grouped in four major categories viz., supervised learning, semi supervised learning, weakly supervised learning and unsupervised learning. A variety of techniques for classification are available in literature including the k-nearest neighbour classifier<sup>2</sup>, Bayesian networks<sup>3</sup>, artificial neural networks<sup>4</sup> and decision trees<sup>5</sup>. Neural networks are one of the most used classification techniques<sup>6</sup>, but they are sensitive to the presence of noise in training data<sup>7</sup>.

<sup>1</sup>Agricultural Research Station (SKNAU, Jobner), Fatehpur-Shekhawati, Sikar 332301, India. <sup>2</sup>Department of Mathematics and Statistics, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana 125004, India. <sup>3</sup>Department of Genetics and Plant Breeding, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana 125004, India. <sup>4</sup>Department of Seed Science and Technology, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana 125004, India. <sup>5</sup>ICAR-Indian Institute of Wheat and Barley, Karnal, Haryana 132001, India. <sup>6</sup>Department of Irrigation and Drainage Engineering, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Udham Singh Nagar, Uttarakhand 263145, India. <sup>7</sup>Department of Pharmaceutics, College of Pharmacy, King Khalid University, 61421 Abha, Asir, Saudi Arabia. <sup>8</sup>Department of Teaching and Learning, College of Education and Human Development, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia. <sup>9</sup>Department of Chemistry, University of Ha'il, 81441 Ha'il, Saudi Arabia. <sup>10</sup>School of Electronic Engineering, Kyonggi University, Yeongtong-gu, Suwon, Gyeonggi-do 16227, Republic of Korea. <sup>11</sup>Department of Environmental Science, Parul Institute of Applied Sciences, Parul University, Vadodara, Gujarat 391760, India. <sup>12</sup>Environmental and Atmospheric Sciences Research Group, Scientific Research Center, Al-Ayen University, Nasiriyah, Thi-Qar 64001, Iraq. ✉email: dinesh.vishwakarma4820@gmail.com; choiardor@hanmail.net

The above classification algorithms for two class and multi class problems have done well for many data sets and for many years. In past few decades, support vector machines (SVM) has emerged as one of the best classification techniques<sup>8</sup>. Vapnik<sup>9</sup> introduced SVM as a machine learning model which is based on kernels for regression and classification task. SVM competes with the performance of other classification techniques and in many cases has shown that it is superior to them for solving classification problems<sup>10</sup>. The SVM is a nonparametric algorithm and is attaining the popularity due to its promising performance and many attractive features. SVM is generally based on the Vapnik Chervonenkis (VC) dimension and the principle of risk minimization. In comparison to neural networks, SVMs are experiencing significant improvements in generalization ability and overcoming other problems like the curse of dimensionality and local minima<sup>9</sup>, which gives SVMs quite a strong competitive advantage over other methods. Also, the principle behind SVM is to find the optimum decision boundary that separates the classes by maximizing the margin. Support vectors are the training data points closest to this maximum margin hyperplane.

The term "Weighted Accuracy Ensemble (EWA)" in the context of SVM refers to an ensemble method that combines multiple models by weighting them according to their accuracy<sup>11,12</sup>. This approach aims to improve the overall prediction performance by giving more importance to models that perform better<sup>13</sup>. In summary, the EWA approach in SVM involves training multiple models, assigning weights based on their accuracy, and combining their predictions to improve the overall model performance<sup>14</sup>.

Particle Swarm Optimization (PSO) techniques have been effectively applied to optimize SVM classifiers for multi-class classification tasks. Various studies have demonstrated the benefits of using PSO to enhance the performance of SVM in complex real-world problems<sup>15–20</sup>. PSO helps in tuning the parameters of SVM, such as regularization and kernel parameters, leading to improved classification accuracy and generalization performance<sup>21–23</sup>. Additionally, the use of PSO with time-varying acceleration coefficients has shown faster convergence and higher precision in optimizing SVM parameters for fault diagnosis applications<sup>24</sup>. The integration of PSO with SVM in multi-domain fusion scenarios, such as corn kernel collision sound signal recognition, has also proven to achieve higher recognition rates for different kernel types<sup>25</sup>.

In recent times, there has been a proliferation of machine learning techniques being explored to identify specific types of wheat<sup>26–35</sup>. Ardjani et al.<sup>15</sup> proposed the use of PSO to optimize the performance of the SVM classifier in multiclass classification problems on the TIMIT corpus, and the results showed that the PSO-SVM approach achieved better classification accuracy compared to other methods, despite an increase in execution time. Luo et al.<sup>20</sup> proposed a novel approach for constructing multi-class least squares wavelet SVM (LS-WSVM) classifiers using quantum particle swarm optimization algorithm (QPSO). The approach optimizes the regularization parameters and kernel parameters of LS-WSVM using QPSO, resulting in improved LS-WSVM models for multi-class classification. The result demonstrates the effectiveness of the approach by conducting simulations, which shows that the proposed method can obtain optimal parameters for LS-WSVM with global searching QPSO to achieve excellent precision for classification. Huang and Dun<sup>17</sup> proposed PSO-SVM model achieved high classification accuracy by simultaneously optimizing the input feature subset selection and the SVM kernel parameter setting. Experimental results showed that the proposed approach correctly selected discriminating input features, and achieved high classification accuracy. Dudzik et al.<sup>18</sup> proposed evolutionary technique optimizes critical aspects of SVMs, including the training sample, kernel functions and features further improving performance in binary classification tasks. Extensive experimental study conducted over more than 120 benchmarks showed that the proposed algorithm outperforms popular supervised learners and other techniques for optimizing SVMs reported in the literature. Hitam et al.<sup>19</sup> optimized SVM model based on PSO for cryptocurrency forecasting and demonstrated that an optimized SVM-PSO algorithm enables accurate prediction of future cryptocurrency prices, surpassing the performance of individual SVM algorithms. Nugraha et al.<sup>24</sup> implemented the PSO-SVM algorithm to classify international journals using the SCImago Journal Rank (SJR) dataset. The accuracy results obtained from PSO-SVM using Linear kernels were 63.12%. Based on these results, PSO-SVM is still unable to optimize the approach in the SJR classification system to achieve 100% accuracy. Sheela and Arun<sup>25</sup> proposed hybrid PSO-SVM algorithm achieved a specificity of 0.85, a sensitivity of 0.956 and an accuracy of 95.78% in determining the presence of pneumonia due to COVID-19.

Proposed methods have been widely used to classify subjects and train efficient ML models in various fields. However, the implication of SVM in agriculture, in particular classifying wheat genotypes, is so far limited. SVM can help predict grain yield by utilising early accessible information on highly heritable and correlated traits. Such ML models may focus on helping plant breeders in early selection and improving selection accuracy to enhance overall genetic gain, thereby ensuring food security. To the best of our knowledge, SVM has been utilised for the very first time to classify the wheat crop based on the grain yield and to train ML models to predict those classes utilising correlated agronomic traits. The integration of ensemble algorithm and particle swarm optimized in support vector machines for classification of wheat genotypes introduces a novel avenue, promising accurate classification of wheat genotype. With the underwritten objective, SVM with six different kernels (linear, radial basis function, sigmoid, polynomial degree-1, degree-2, and degree-3) were employed to classify 302 wheat (*Triticum aestivum*) genotypes belonging to different classes of breeding material viz. improved genotypes, landraces, varieties, and advanced breeding lines:

- (a) To select the suitable kernels (Linear, RBF, Sigmoid and Polynomial with degree 1, 2 and 3) for training Support Vector Machines which classify the 302 wheat genotypes most accurately using 14 morphological attributes.
- (b) The EWA approach is proposed by combining the outputs of individual classifiers with six kernel functions. The accuracy was used in the ensemble weighting process to enhance the classification of wheat genotypes.

- This has the potential to significantly enhance the SVM classification model's ability for learning and generalization.
- (c) To optimize the SVM hyper-parameters of best kernel out of the six studied kernels for classification of wheat genotypes using GS, RS, GA, DE and PSO techniques. By doing so, the SVM's parameters are optimized, which directly influencing the efficiency and predictive power of SVM, which ultimately enhancing classification accuracy.

To compare the performance of various kernels, EWA approach and PSO optimization techniques used. With appropriate classification technique, the low, medium and high yielding wheat genotypes can be discovered accurately. The rest of paper is organized as follows. Section "Support vector machine (SVM) algorithms" introduces the support vector machine (SVM) algorithms and the related work presented in this section. The Materials and Methods is presented in "Materials and Methods" section. In "Results and Discussion" section, the Comparative performance of various kernels and ensemble approach for SVM classification of wheat genotypes and Comparative performance of optimization approaches for SVM classification of wheat genotypes is discussed with objective evaluation measures. Finally, a conclusion is touched and future work is discussed in "Conclusions and future research" section.

### Support vector machine (SVM) algorithms

SVMs have been used in regression and classification problems<sup>36</sup>, and in these two different cases SVMs are called support vector regression (SVR) and support vector classifier (SVC), respectively<sup>37</sup>. Only SVC has been included here and is referred as 'SVM' uniformly throughout the current manuscript. There are two types of SVMs: 1) nonlinear SVMs and 2) linear SVMs, depending on whether or not the data needs to be transformed into higher dimension, respectively<sup>38</sup>. Linear SVMs can be further categorized into two groups, linear SVM for separable and non-separable cases. A linear separable hyperplane can be designed for classification of all training data points without any misclassification error in former. While in latter case, a linear separable hyperplane exists at the expense of some training errors. A kernel function serves as a bridge between these linearly separable and non-separable data. Some of the most important kernel functions are.

- (i) Linear kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)$ ,
- (ii) RBF kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp[-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2]$ , where  $\gamma$  is known as the gamma/sigma parameter,
- (iii) Sigmoid kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh[(\mathbf{x}_i^T \mathbf{x}_j) + b]$ , where  $b$  is the parameter and
- (iv) Polynomial kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = [\gamma (\mathbf{x}_i^T \mathbf{x}_j) + 1]^d$ , where  $d$  and  $\gamma$  are the degree and scale parameters, respectively.

In SVMs, we are exploring two things: 1) a hyperplane with maximum margin and 2) a hyperplane with minimum misclassification rate. The parameter  $C$  is a cost parameter which controls the agreement between maximum margin and minimum misclassification rate<sup>39</sup>. A low value of  $C$  allows more outliers, whereas a high value of  $C$  allows fewer outliers. The simplest kernel function is linear kernel and is given by the inner product. The parameter  $\gamma$  in RBF and polynomial kernels determines the reach of a training point. If the value of  $\gamma$  is high, the decision boundary of SVM will depend on training points that are closest to the decision boundary by ignoring the training points that are farther<sup>40</sup>. While, a low  $\gamma$  value would result in a decision boundary that would consider training points that are far from it. Therefore, higher  $\gamma$  values usually produce highly flexible decision boundaries and lower  $\gamma$  values lead to decision boundaries that are more linear.

SVM was initially designed to classify linearly separable and non-separable data, and later expanded to a non-linear domain through the use of kernel functions<sup>41</sup>. A kernel function serves as a connector between these linearly separable and non-separable data. One of the problems with the use of SVM methodology is to select a kernel function which depends upon the task and dataset. There is no consensus as to which kernel is better or worse for specified applications.

### Kernel function and optimization techniques in SVM

The most commonly used kernel functions are linear, Radial Basis Function (RBF), sigmoid and polynomial kernels<sup>42</sup>. Therefore, in first phase of this study, a comparative analysis was carried out between linear, RBF, sigmoid and polynomial (with degree 1, 2 and 3) kernels with the objective to find the kernel which classified wheat genotypes most accurately. In this phase, the outputs of individual classifiers with six kernel functions were also combined using an Ensemble with Weighted Accuracy approach<sup>43</sup>. The accuracy was used in the ensemble weighting process to enhance the classification of genotypes.

A good set of SVM parameters plays the critical role to improve its classification performance in addition to kernel selection. Various studies have been conducted to select these parameters, but there is no general view of their settings<sup>44</sup>. To overcome this problem, several deterministic and probabilistic algorithms have been considered to optimize the SVM and kernel parameters for classification. Grid Search (GS) and Random Search (RS) are most used deterministic algorithms, because of their good results and simplicity. GS and RS are not statistically feasible for optimizing the hyper parameters in large datasets. In such situations, probabilistic optimization algorithms like Genetic Algorithm (GA), Differential Evolution (DE) and Particle Swarm Optimization (PSO) are generally used<sup>45,46</sup>. These five optimization algorithms constitute the second phase of this study to optimize SVM hyper-parameters for classification of wheat genotypes.

### Multiclass classifier

Initially, SVM was proposed for performing binary classification task. However, binary classification applications are very limited. Several methods have been proposed by researchers to generate multiclass SVM from binary SVM and are still an ongoing research topics. To this end, *one-versus-all* (*one against the rest*) technique was developed to solve multiclass problems. Vapnik<sup>47</sup> recommended grouping one class with others. In *one against the rest*, separate binary classifiers are trained to discriminate one class from the rest of classes<sup>48</sup>. For  $k$  classes,  $k$  quadratic programming problems must be solved. Another way is the *one-versus-one* or *pair wise comparison*<sup>49</sup>. There are  $k(k-1)/2$  binary classifiers for every possible pair of  $k$  classes. In this approach, classifier counts are typically much larger than *one against the rest*, whereas the number of training observations required for each classifier is very small. Therefore, this approach is treated to be more efficient than that of the *one against the rest* method.

Pal<sup>50</sup> suggested the suitability of *one-versus-one* approach when comparing the performance of six multiclass approaches (*one-versus-one*, *one-versus-all*, DAG, Error Corrected Output Coding, Bound constrained and Cramer & Singer) to solve classification problem in term of computational cost and classification accuracy. Therefore, in this study we employed *one-versus-one* methodology for multiclass problem.

### Ensemble SVM approach

The performance of learning algorithms with six kernel functions depends on various model configurations, such as model parameters and input feature types<sup>51</sup>. To tackle the limitation of individual model performance, the wheat dataset was further classified with an SVM based ensemble learning algorithm to improve the training and testing accuracy. Ensemble learning was proposed for reducing classification bias and variance. Ensemble learning is considered as one of the most effective strategies to balance the influence from bias and variance in classification tasks<sup>52</sup>. The ensemble methods aggregate different algorithms together for a comprehensive decision. There were several ensemble learning approaches proposed in the literature, such as: majority voting, averaging, weighted averaging, stacking, bagging and boosting.

For the multi-class wheat dataset, we proposed an Ensemble with Weighted Accuracy approach in which accuracy was used for the ensemble weighting process to improve the classification of genotypes. Given an input feature vector  $\mathbf{x} \in R^p$ , a classification result  $g_m(\mathbf{x}) \in y_i$  is obtained based on each base classifier  $g_m \in B$ , where  $B$  is the classifier set  $B = \{g_m : m = 1, 2, \dots, t\}$ , where  $t = 6$  is the number of base classifiers. The aim of the ensemble learning algorithm is to create an improved composite classifier  $E(\mathbf{x})$ , by amalgamating the classification outputs from the different base classifiers into an improved output. The accuracy was used to weigh different classification outputs of the base classifiers in the EWA approach. The class variables for wheat genotypes were labelled as 1 (for L class), 2 (for M class) and 3 (for H class). The concept behind this approach for wheat genotypes was:

$$E(\mathbf{x}) = \begin{cases} 1 & 1 \leq \sum_{m=1}^t w_m g_m(\mathbf{x}) < 1.67 \\ 2 & 1.67 \leq \sum_{m=1}^t w_m g_m(\mathbf{x}) < 2.33 \\ 3 & 2.33 \leq \sum_{m=1}^t w_m g_m(\mathbf{x}) \leq 3 \end{cases} \quad (1)$$

where,  $w_m$  is the weight for each base classifier. The training and testing accuracies were used to weigh the classification results of training and testing datasets, respectively, as:

$$w_m = \frac{Acc_m}{\sum_{j=1}^t Acc_j}, \forall m \quad (2)$$

where,  $Acc_m$  is the accuracy of  $m$ -th base classifier.

### Optimization algorithms for SVMs

In general, most of the machine learning approaches will not generate optimal results if their hyper-parameters are not properly adjusted. Parameter tuning can be time-consuming especially when done manually for the learning algorithms with multiple parameters<sup>53</sup>. To overcome, a few SVM optimization methods have been suggested to address the optimization of kernel parameters. GS is the most prevalent deterministic algorithm to determine the appropriate values of parameters<sup>54</sup>. The parameter values leading to the highest classification accuracy can be obtained by setting the appropriate values of the lower and upper limits (search interval) and step length. RS replaces the exhaustive enumeration of all combinations through randomization<sup>55</sup>. To implement RS, the parameter grid was set to utilize random combinations to train the model.

The genetic algorithm, first developed by John Holland in 1975, is a way of solving the problems of optimization in terms of natural selection, a process that drives evolution<sup>56</sup>. GA is a probabilistic search and optimization method that seeks to mimic biological evolution as a problem-solving strategy<sup>57</sup>. To carry out its optimization, GA uses three operators (selection, crossover and mutation) to spread its population from one generation to the next. GA was developed to reduce the training time by using minimum features and to enhance the classification accuracy<sup>39,58</sup>. Differential evolution (DE) is a type of real number coding optimization technique based on population evolution<sup>59</sup>. DE also uses crossover, mutation and selection operators like GA. As crossover is a leading evolutionary strategy for GA, DE considers mutation to be the most important operator. DE rapidly converges and can provide with the optimal solution in most of the situations<sup>60</sup>. In addition, it has proved to be more efficient and powerful in contrast to other optimization approaches<sup>61</sup>.

### Particle swarm optimization algorithm

Kennedy and Eberhart<sup>62</sup> proposed PSO algorithm as a nature-induced metaheuristic approach. The algorithm proposes to mimic fish schooling or the behaviour of a flock of birds is an evolutionary approach and has been used to address many of the sophisticated problems of optimization. As a computational intelligence approach, since it requires fewer tuning parameters, it offers several advantages over other approaches such as robustness, flexibility and higher computational efficiency<sup>63</sup>. There are no mutation and/or crossover operators in PSO in comparison to GA and DE.

Individuals (*particles*) are transported in PSO through a hyper-dimensional search space with a tendency to mimic the success of others in a population (*swarm*). Particles change their positions and velocities within the swarm are greatly influenced by the experience and knowledge of their neighbours. Particles within the search space track their coordinates and are directly related to the best fitness they have gained so far called as personal best (*Pbest*). The overall best value is called global best (*Gbest*). The concept of PSO focuses on changing the velocity of each particle after every iteration as per its *Gbest* and *Pbest*<sup>64</sup>.

Each particle of the swarm is considered a potential candidate for the problem in PSO algorithm. Consider a swarm with the entire particles in the  $p$ -dimensional target search space. In  $j^{\text{th}}$  iteration, the position and velocity of  $i^{\text{th}}$  particle are denoted by two vectors, position vector  $\{x_{i1}^j, x_{i2}^j, \dots, x_{ip}^j\}$  and the velocity vector  $\{v_{i1}^j, v_{i2}^j, \dots, v_{ip}^j\}$ , respectively. To obtain a global optimum and in the process of iteration, velocity of each particle is updated as per its *Pbest* and *Gbest*. The Eqs. (1) and (2) describe how PSO algorithm updates its velocity and position of their particles<sup>65</sup>.

$$v_{ip}^j = wv_{ip}^{j-1} + c_1r_1(Pbest_{ip}^{j-1} - x_{ip}^{j-1}) + c_2r_2(Gbest_p^{j-1} - x_{ip}^{j-1}) \quad (3)$$

$$x_{ip}^j = x_{ip}^{j-1} + v_{ip}^j \quad (4)$$

where,  $j$  is the current iteration number;  $i$  denotes the particle number;  $p$  represents the dimension of feature space;  $v$  is the velocity and  $x$  is the position of each particle;  $v_{ip}^j$  and  $x_{ip}^j$  denotes the velocity and position of  $i^{\text{th}}$  particle in  $p$ -dimensional feature space after  $j^{\text{th}}$  iteration, respectively;  $w$  is the inertia weight or weighting coefficient;  $c_1$  and  $c_2$  are the acceleration coefficients known as social and cognitive parameters, and generally  $c_1 = c_2 = 2$ ;  $r_1$  and  $r_2$  are the random numbers generated with uniformly distribution over the interval  $[0,1]$ ;  $Pbest_{ip}^{j-1}$  is the best position of  $i^{\text{th}}$  particle; and  $Gbest_p^{j-1}$  denotes the best position taken from the swarm.

At  $j^{\text{th}}$  iteration, *Pbest* and *Gbest* of each particle are updated as follows:

$$\text{If } f(x_{ip}^j) < f(Pbest_{ip}^{j-1}) \text{ then } Pbest_{ip}^j = Pbest_{ip}^{j-1} \text{ else } Pbest_{ip}^j = x_{ip}^j \quad (5)$$

$$\text{If } f(x_{ip}^j) < f(Gbest_p^{j-1}) \text{ then } Gbest_p^j = Gbest_p^{j-1} \text{ else } Gbest_p^j = x_{ip}^j \quad (6)$$

where,  $f(x)$  is the fitness function subject to maximization. The updating process should be repeated until it reaches a stop condition, such that a predefined number of iteration is met<sup>66</sup>. A detailed procedure for evaluating SVM parameters using the PSO algorithm is shown in Fig. 1.

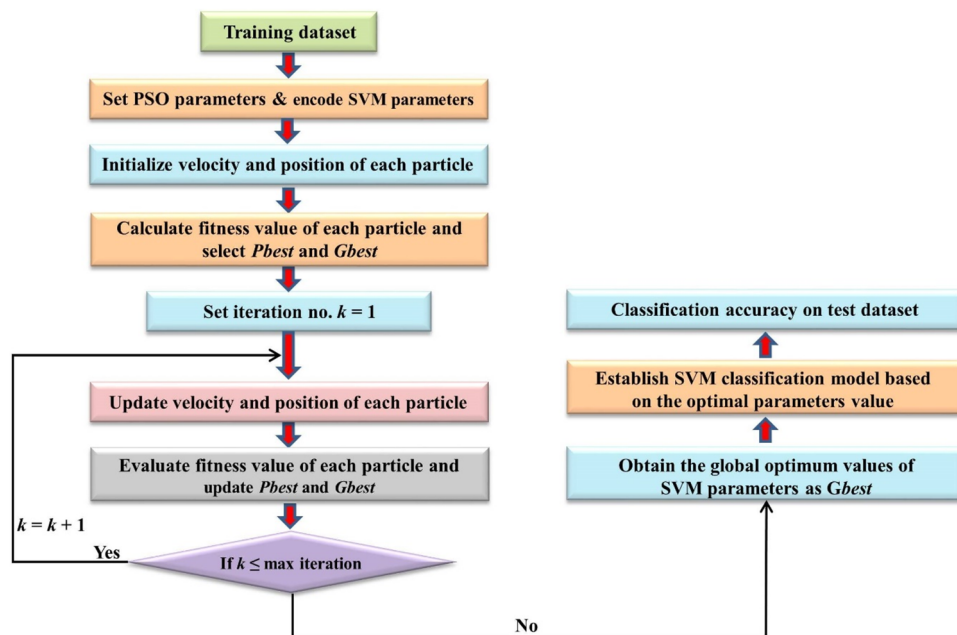
### Related works

The literature has described a number of grading systems that use various morphological aspects for the classification of many cultivars and cereal grains<sup>67,68</sup>. Zhang et al.<sup>69</sup> differentiated the fungal infected and healthy wheat kernels. A multi-class SVM with RBF kernel was used for classification. The wheat kernels infected by *Aspergillus niger*, *Aspergillus glaucus*, and *Penicillium* spp. and healthy wheat kernels were classified with accuracies of 92.9%, 87.2%, 99.3%, and 100%, respectively. Yao et al.<sup>70</sup> presented an application of image processing techniques and SVM for detecting rice diseases. The results showed that SVM effectively detected and classified these disease spots to an accuracy of 97.2%. Jian and Wei<sup>71</sup> recognized the cucumber leaf diseases using RBF, polynomial and sigmoid kernel function. The results showed that the SVM method based on RBF kernel function and taking each spot as a sample made the best performance for classification of cucumber leaf diseases.

Dubey and Jalal<sup>72</sup> experimentally validated a solution for the detection and classification of apple fruit diseases using a multi-class SVM. The classification accuracy for the proposed solution that was achieved was up to 93%. Sengupta and Lee<sup>73</sup> used a novel technique to detect immature green citrus fruit in tree canopy under natural outdoor conditions. The approach was able to accurately detect and count 80.4% of citrus fruit in a validation set of images acquired from a citrus grove under natural outdoor conditions. Bhange and Hingoliwala<sup>74</sup> developed a web-based tool that helps farmers for identifying fruit disease by uploading fruit image to the system. SVM was used for classification to classify the image as infected or non-infected. Experimental evaluation of the proposed approach was effective and 82% accurate to identify pomegranate disease.

Chung et al.<sup>75</sup> proposed an approach to distinguish infected and healthy seedlings of the rice cultivars Tainan 11 and Toyonishiki. SVM classifiers were developed for distinguishing the healthy and infected seedlings. GA was used for selecting essential traits and optimal model parameters for the SVM classifiers. The proposed approach distinguished healthy and infected seedlings with a positive predictive value of 91.8% and an accuracy of 87.9%. Padol and Yadav<sup>76</sup> intended to aid in the detection and classification of grape leaf diseases using SVM





**Fig. 1.** Flowchart for SVM parameter optimization using PSO algorithm.

classification technique. The proposed system can successfully detect and classify the examined disease with 88.89% accuracy. Bonah et al.<sup>66</sup> presented a classification model of SVM together with GS, GA and PSO optimization algorithms for bacterial foodborne pathogen classification. Simulated results show training accuracies of 100% and prediction accuracies of 98.95% for five selected bacterial pathogens acquired using electronic nose dataset and PSO-SVM model.

## Materials and methods

### Data acquisition

The data used in the study consists of 302 genetically diverse bread wheat (*Triticum aestivum*) genotypes belonging to different classes of breeding material viz. improved genotypes, landraces, varieties, and advanced breeding lines (Supplementary material Table S1). The secondary data were taken from an experiment conducted during Rabi season of 2018–19 by the Department of Genetics and Plant Breeding, Chaudhary Charan Singh Haryana Agricultural University, Hisar (latitude 29° 09' 6.70" N, longitude 75° 43' 16.04" E and altitude 215 m). The site of the experiment comes under sub-tropical to semi-arid zone. The experiment was conducted in alpha lattice design with two replications and the seed material was sown in the first week of November-2018 using a dibbler called as IIWBR dibbler<sup>77</sup>. Each genotype occupied a plot size of 1 m<sup>2</sup> (1.25 m × 0.8 m). Row to row (plant to plant) spacing was 20 cm (10 cm) with seedling depth of 5 cm. Recommended package of practices were used to raise a genotype and a population of thirty plants were maintained throughout the trial. In each plot, after leaving the border lines, ten random plants were selected and tagged to obtain the data on following 14 morphological attributes viz., Days to heading (DTH), Days to anthesis (DTA), Days to maturity (DTM), Grain filling duration (GFD), Number of tillers/plant (NTP), Plant height (PH), Peduncle length (PL), Spike length (SL), Spikelets/spike (SS), Total number of grains/plant (TNG), Thousand kernel weight (TKW), Grain yield/plant (GY), Biomass/plant (BM) and Harvest index (HI).

### Pre-processing of data

Initially, the correlation plots were constructed for wheat dataset to check the association between different morphological variables. Correlation between variables clearly depicted that there was very low amount of linear association among various variables and only four variables (TNG, TKW, BM and HI) were significantly correlated with grain yield (Fig. 2). The characteristic grain yield (g/plant) was used as class attribute. The Jenks natural breaks optimization method was utilized to change the continuous attribute into class attribute<sup>78</sup>. It is a data clustering method developed to find the best combination of values into various classes. The method attempts to minimize the within class variance and to maximize the between class variance. The class attribute grain yield was categorized into 3 classes i.e., low yield (8.70–14.68 g/plant), medium yield (14.69–18.05 g/plant) and high yield (18.06–24.04 g/plant). Out of 302 genotypes, the Jenks method classified 102, 128 and 72 genotypes in low (L), medium (M) and high (H) yield classes, respectively (Fig. 3). Out of the 302 wheat genotypes, the number of genotypes in training (and testing) dataset were 82 (20), 103 (25) and 58 (14) for L, M and H yield classes, respectively. Descriptive statistics of the morphological attributes with respect to individual classes as well as overall are given in Table 1. The minimum and maximum of mean value of genotype were reported 10.70 (SD ± 1.30), 10.50 (SD ± 1.20), 10.40 (SD ± 1.30) and 10.50 (SD ± 1.30); and 329.10 (SD ± 37.90), 384.10 (SD ± 41.60), 435.80

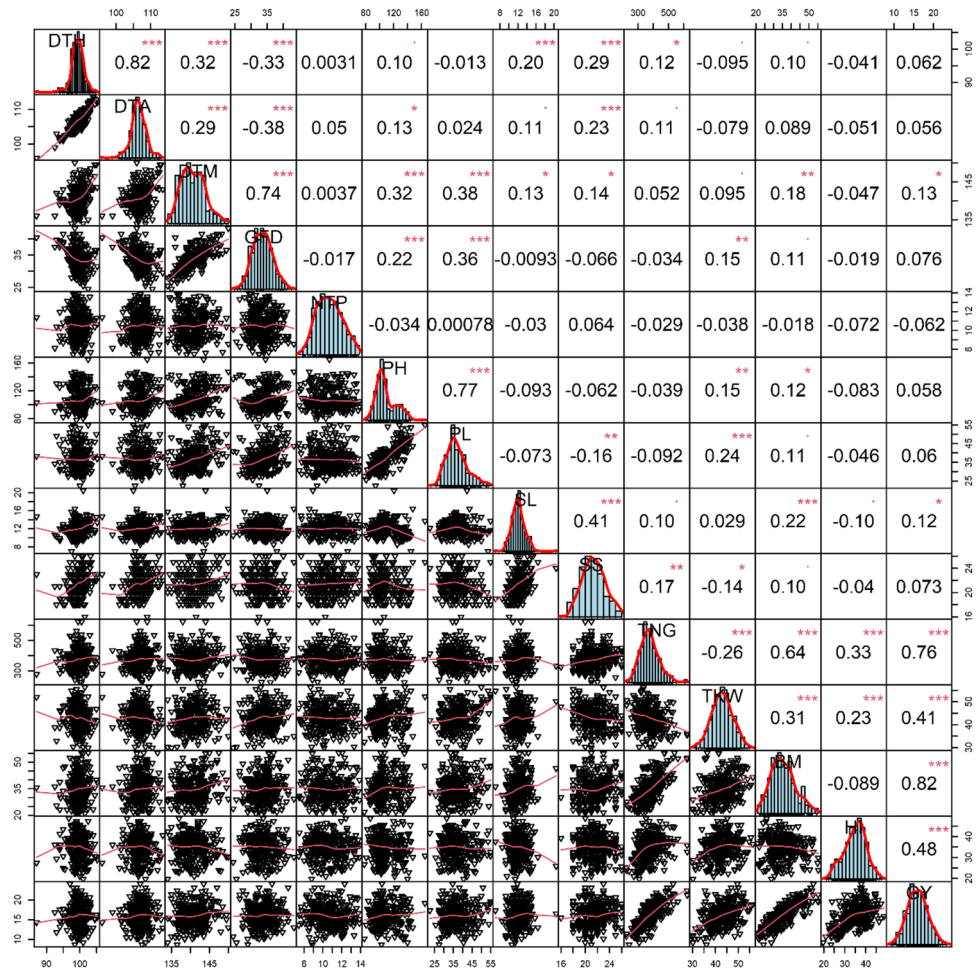


Fig. 2. Scatter plot showing correlation among various wheat variables.

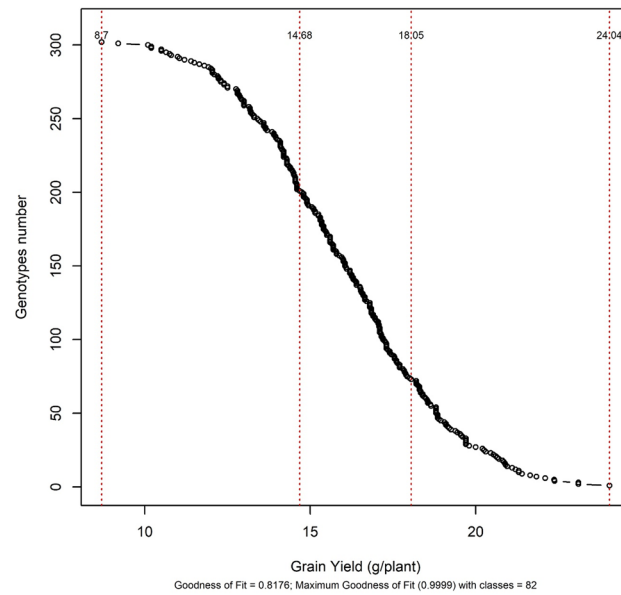


Fig. 3. Jenks natural breaks optimization technique classifies genotypes into three classes.

Variables	Low yield (102 genotypes)			Medium yield (128 genotypes)			High yield (72 genotypes)			Overall (302 genotypes)		
	Mean	SD	CV	Mean	SD	CV	Mean	SD	CV	Mean	SD	CV
DTH	99.3	1.6	1.6	99.7	1.7	1.7	99.7	1.4	1.4	99.6	1.6	1.6
DTA	106.6	2.0	1.9	106.8	1.9	1.8	107.2	1.7	1.6	106.8	1.9	1.8
DTM	140.4	2.5	1.8	140.9	2.9	2.1	141.1	3.1	2.2	140.8	2.9	2.1
GFD	33.6	3.0	8.9	33.8	3.0	8.9	33.7	3.0	8.9	33.7	3.0	8.9
NTP	10.7	1.4	13.1	10.5	1.5	14.3	10.4	1.4	13.5	10.5	1.4	13.3
PH	108.5	15.2	14.0	105.7	12.4	11.7	112.5	15.1	13.4	108.3	14.3	13.2
PL	37.0	5.6	15.1	36.3	4.8	13.2	38.0	5.9	15.5	36.9	5.4	14.6
SL	11.8	1.3	11.0	12.2	1.2	9.8	12.2	1.3	10.7	12.1	1.3	10.7
SS	21.1	1.9	9.0	21.6	1.8	8.3	21.7	2.0	9.2	21.5	1.9	8.8
TNG	329.1	37.9	11.5	384.1	41.6	10.8	435.8	44.8	10.3	377.8	57.6	15.2
TKW	40.7	5.1	12.5	43.2	4.3	10.0	45.5	4.9	10.8	42.9	5.1	11.9
BM	29.2	3.9	13.4	35.4	4.4	12.4	43.0	4.4	10.2	35.1	6.7	19.1
HI	31.7	5.7	18.0	35.9	4.5	12.5	37.1	4.4	11.9	34.8	5.4	15.5

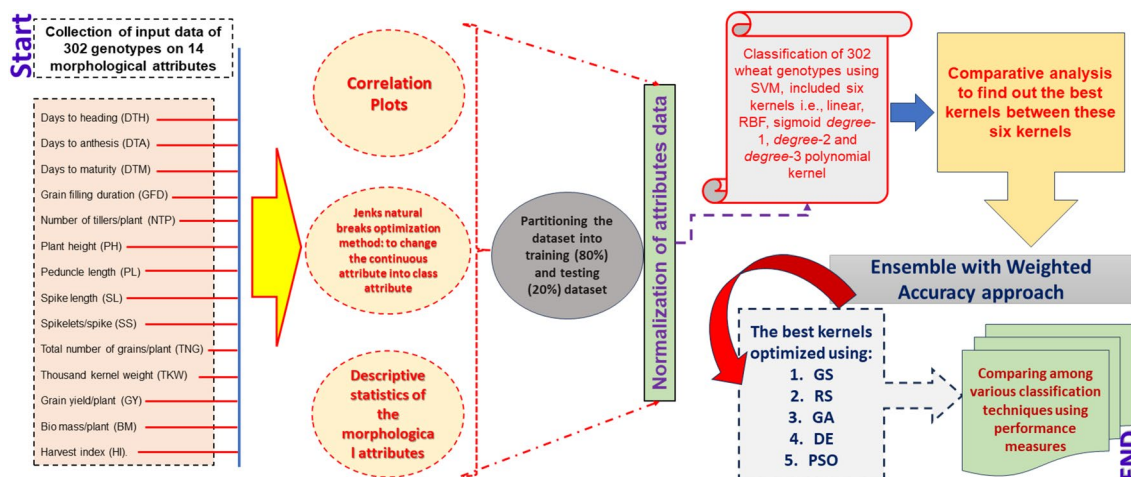
**Table 1.** Class-wise and overall descriptive statistics for wheat variables.

(SD  $\pm$  44.80) and 377.80 (SD  $\pm$  57.60), respectively for low yield, medium yield, high yield, and overall wheat variables. While the minimum coefficient of variance was found 1.60, 1.70, 1.40 and 1.60, respectively for low yield, medium yield, high yield, and overall wheat variables and maximum value 18.00, 14.30, 15.50 and 19.10, respectively. The minimum (maximum) value was reported in NTP (TNG) variable for low yield, medium yield, high yield and overall wheat variables.

For the model development, it is essential to divide the dataset into training dataset and test dataset<sup>66</sup>. Model training is performed using the training dataset and test datasets are used for the performance assessment of the classification model. Therefore, the dataset was partitioned into training (80%) and test (20%) dataset with stratified sampling method. 10 fold repeated cross validation with five repeats was used as resampling method for partitioning the training and test datasets<sup>60</sup>. Prior to partitioning, the whole dataset was normalized to zero mean and unit variances to reduce the learning time and improve the performance of the model.<sup>55</sup>.

### Proposed methodology

Figure 4 shows the proposed methodology diagram of the study. For the classification of wheat genotypes using SVM, this study included six kernels i.e., linear, RBF, sigmoid *degree-1*, *degree-2* and *degree-3* polynomial kernel. Initially, a comparative analysis was carried out between these six kernels with the objective to find the kernel which classified wheat genotypes most accurately. Then, the outputs of individual classifiers with six kernel functions were combined using an Ensemble with Weighted Accuracy approach where the accuracy was used in the ensemble weighting process to enhance the classification of genotypes. The best among these kernels was further taken forward for optimization based SVM classification. Then the performance of PSO-SVM was compared with GS, RS, GA and DE optimized SVM classification by optimizing the parameters. Finally, the one-way ANOVA was conducted separately for training and testing datasets to assess the difference in the accuracy of all twelve classifiers is statistically significant or not<sup>79</sup>. The Tukey HSD post hoc test was performed, which makes



**Fig. 4.** shows the proposed methodology diagram.



pairwise comparisons of accuracies of classifiers to find out whether their difference is significant at the desired significance level (0.05 in this study). The application platform was implemented in R version 4.0.2. The BAM-Mtools, classInt, Libsvm, e1071, caret (Classification and Regression Training), GA, DEoptim and pso packages were used for SVM classification and optimization.

To achieve the proposed methodology, we used a workstation which was configured with an Intel Xenon processor of 24 cores (48 threads), 64 GB RAM, 1 TB SSD and 4 GB NVIDIA graphic card. Each analysis was done in parallelization activating  $n-2$  cores.

### Performance measures

The study comprises of widely used performance measures such as accuracy, Kappa value, sensitivity, specificity, predictive values, balanced accuracy and F-measure<sup>80</sup>. Using the confusion matrix  $C = (C_{ij})_{3 \times 3}$ , the following performance measures have been used in this study:

#### Accuracy

It is one of the most widely used measures in classification performance, and is defined as the ratio between correctly classified observations and the total number of observations<sup>81</sup>.

$$Accuracy = \frac{C_{11} + C_{22} + C_{33}}{\sum \sum C_{ij}} \quad (7)$$

#### Kappa value

It estimates how well the observations classified by the classifier match the observations labelled as grounded. It is considered a more robust measure than a simple per cent agreement calculation, as it considers the probability of agreement that occurs by chance<sup>82</sup>. Thus, Kappa is generally slightly lower than accuracy<sup>83</sup>.

$$Kappa = \frac{p_o - p_e}{1 - p_e} \quad (8)$$

where,  $p_o$  is the overall accuracy and  $p_e$  is calculated as:

$$p_e = \frac{\sum_{i=1}^3 R_i C_i}{(\sum R_i)(\sum C_i)} \quad (9)$$

where,  $R_i$  and  $C_i$  are the number of predicted and actual observations for  $i^{th}$  class, respectively.

#### Sensitivity and specificity

Sensitivity is also known as True Positive Rate (TPR) or Recall. It represents the proportion of correctly predicting the selected class<sup>81</sup>. Specificity or True Negative Rate (TNR) shows the proportion of correctly predicting the non-selected other classes<sup>81</sup>.

$$Sensitivity \text{ or } Recall = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP} \quad (10)$$

Suppose there are three classes in this study viz., L, M and H; then: 1) True Positive (TP) are all L class instances that are classified as L, 2) True Negative (TN) are all non-L class instances that are not classified as L or classified as M and H, 3) False Positive (FP) are all non-L class instances that are classified as L, and 4) False Negative (FN) are all L class instances that are not classified as L or classified as M and H.

#### Predictive values

These values represent the predictive performance. Precision or Positive Predictive Value (PPV) represents the ratio of selected class correctly predicted to the total selected class predictions made<sup>81</sup>. Negative Predictive Value (NPV) represents the proportion of the non-selected other classes correctly predicted to the total non-selected other classes predictions made<sup>80</sup>.

$$PPV = \frac{TP}{TP + FP}, \quad NPV = \frac{TN}{TN + FN} \quad (11)$$

#### Balanced accuracy

The average of sensitivity and specificity is known as balanced accuracy<sup>84</sup>.

$$BalancedAccuracy = \frac{1}{2} (Sensitivity + Specificity) \quad (12)$$

#### F-measure

It is the harmonic mean of precision or positive predictive value and sensitivity or recall<sup>81</sup>. The value of F-measure ranges from 0 to 1.

$$F\text{-measure} = \frac{(2 \times \text{Precision}) \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

## Results and discussion

### Comparative performance of various kernels and ensemble approach for SVM classification of wheat genotypes

During the training phase, a 10 fold repeated cross validation with 5 repeats was used for selecting the SVM parameters. First, dataset was randomly divided as the training sets of 10 mutually exclusive subsets of equal size. Then we trained an SVM classifier modeled with predefined parameter values for different kernels, each one 10x. Every time we omitted a subset of the training dataset and used it only to get a measure of classification accuracy. From 10x of training and accuracy calculations, testing accuracy revealed the predictive performance of SVM classifier<sup>85</sup>. We then selected the best parameter values of SVM classifier to maximize the prediction. The procedure for parameter estimation was repeated five times, with each of the five different training sets randomly generated.

The number of support vectors, training and testing accuracies and Kappa values have been summarized in Table 2 for the seven models including the ensemble approach. All the six kernel functions were used with their default parameter settings. The classification accuracies over training (testing) datasets of 93.4% (93.2%), 93.8% (93.2%), 81.5% (86.4%), 93.0% (84.7%), 76.1% (61.0%) and 91.8% (78.0%) were obtained for linear, RBF, sigmoid, degree-1 polynomial, degree-2 polynomial and degree-3 polynomial kernels, respectively. Results indicated that the performance of RBF kernel was better than other kernels in terms of accuracies and Kappa values. Therefore, RBF kernel appears to be effective than the other kernel counterparts. So, we considered RBF kernel for the next phase of the study and optimized its parameter using various optimization algorithms.

The outcomes of RBF kernels were more consistent with that of Chung et al.<sup>75</sup> who distinguished infected and healthy seedlings of rice with 87.9% accuracy. Zhang and Wu<sup>86</sup> demonstrated that the *one-versus-one* multiclass Gaussian kernel SVM can attain 88.2% accuracy. Here the accuracy with RBF kernel were higher than reported by Bulanon et al.<sup>87</sup> (87%) and by Lu et al.<sup>88</sup>(87.2%). Zhang et al.<sup>89</sup> demonstrated that the polynomial kernel can classify a dataset with 88.83% accuracy. Manurung et al.<sup>90</sup> also found the similar results with 81.76% testing accuracy for polynomial kernel when comparing the kernel functions for Australian credit approval data of UCI machine learning repository. Melgani and Bazi<sup>85</sup> confirmed the superiority of Gaussian RBF kernel based SVM classification as compared to linear and polynomial kernels.

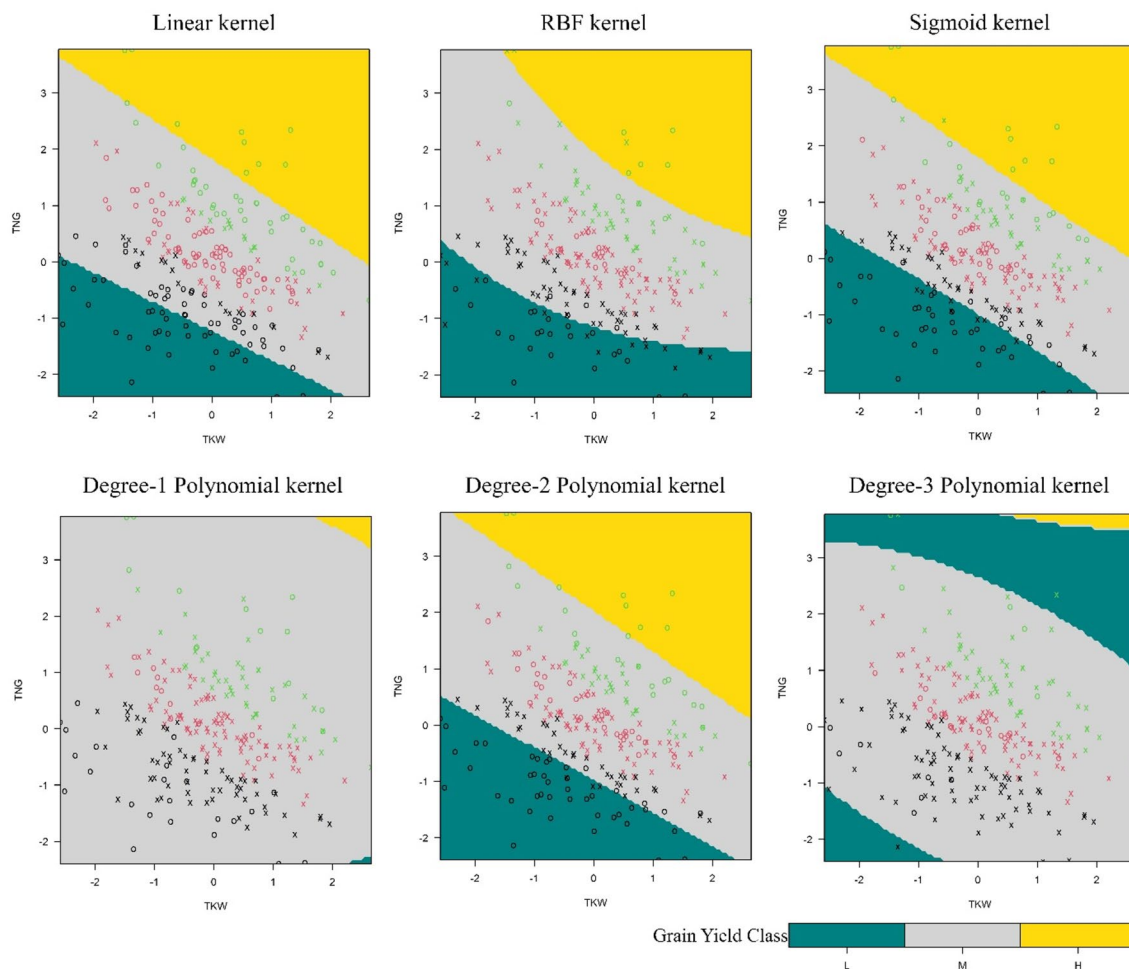
Figure 5 shows the two-dimensional SVM plots of the training data with the separating hyperplane (decision surface) superimposed on the training data for the case of (a) Linear, (b) RBF, (c) Sigmoid, (d) Degree-1 polynomial, (e) Degree-2 polynomial and (f) Degree-3 polynomial kernels. These plots offer a visual representation of the three yield classes' data as a function of any two variables (total number of grains and thousand kernel weight in this case). Examples of misclassification cases can be seen in all the six kernels with individual training data being located on the wrong side of the separating hyperplane.

The comparison of results of each model with the ensemble approach show that the EWA approach structure outperformed others in terms of accuracy and Kappa value. The best classification accuracy over training dataset (Kappa value) achieved by the EWA approach was 95.1% (92.4%). Specially, EWA approach also represents the highest classification accuracy over testing dataset (Kappa value) of 94.9% (92.2%).

Table 3 shows the confusion matrix and various performance measures for the SVM classification of wheat genotypes using linear and RBF kernel functions. It is obvious from this table that both kernels produced similar accuracy of 100% for the high yield class. The table also shows that out of the 59 genotypes (20 for low yield, 25 for medium yield and 14 for high yield), the linear kernel-based classification predicted 18 low yield genotypes and 23 medium yield genotypes accurately, whereas 2 low yield genotypes were misclassified in medium class and 2 medium yield genotypes were mislabelled in low and high classes each with 1 genotype. The RBF kernel classification predicted 19 low yield genotypes and 22 medium yield genotypes accurately while 1 low yield genotype was misclassified in medium class and 3 medium yield genotypes were mislabelled in low (1 genotype) and high (2 genotypes) classes. The average sensitivity, specificity, PPV, NPV, balanced accuracy and F-measures were calculated and are equal to 94.0% (94.3%), 96.4% (96.7%), 93.3% (92.7%), 96.4% (96.4%), 95.2% (95.5%) and 93.6% (93.3%), respectively for linear (RBF) kernel functions.

Kernel	Number of Support vectors	Training		Testing	
		Accuracy	Kappa	Accuracy	Kappa
Linear	67	0.934	0.899	0.932	0.896
RBF	190	0.938	0.905	0.932	0.897
Sigmoid	143	0.815	0.714	0.864	0.790
Degree-1 polynomial	145	0.930	0.892	0.847	0.765
Degree-2 polynomial	211	0.761	0.620	0.610	0.382
Degree-3 polynomial	195	0.918	0.871	0.780	0.648
EWA approach	-	0.951	0.924	0.949	0.922

**Table 2.** Comparative performance of different kernels for SVM-based classification of wheat genotypes.



**Fig. 5.** Two-dimensional SVM plots showing training data and decision boundaries using attributes TNG and TKW for the case of (a) Linear, (b) RBF, (c) Sigmoid, (d) Degree-1, (e) Degree-2 and (f) Degree-3 polynomial kernels.

Performance statistics	Actual					
	Linear kernel			RBF kernel		
	L	M	H	L	M	H
Prediction						
L	18	1	0	19	1	0
M	2	23	0	1	22	0
H	0	1	14	0	2	14
Sensitivity	0.900	0.920	1.000	0.950	0.880	1.000
Specificity	0.974	0.941	0.978	0.974	0.971	0.956
Positive predictive value	0.947	0.920	0.933	0.950	0.957	0.875
Negative predictive value	0.950	0.941	1.000	0.974	0.917	1.000
Balanced accuracy	0.937	0.931	0.989	0.962	0.925	0.978
F-measure	0.923	0.920	0.966	0.950	0.917	0.933

**Table 3.** Confusion matrix and performance measures for SVM classification of wheat genotypes using Linear and RBF kernels.

The confusion matrix obtained by sigmoid and degree-1 polynomial kernels is shown in Table 4. It shows that low yield class was correctly classified for 19 genotypes and 3 medium yield genotypes were misclassified as low yield, in case of sigmoid kernel. Whereas, medium yield class was correctly classified for 22 genotypes and high yield class was correctly classified for 11 genotypes. For degree-1 polynomial kernels, 16 low yield genotypes, 21 medium yield genotypes and 13 high yield genotypes were accurately classified whereas 4 low yield genotypes

Performance statistics	Actual					
	Sigmoid kernel			Degree-1 polynomial kernel		
	L	M	H	L	M	H
Prediction						
L	19	3	0	16	2	0
M	1	21	3	4	21	1
H	0	1	11	0	2	13
Sensitivity	0.950	0.840	0.786	0.800	0.840	0.929
Specificity	0.923	0.882	0.978	0.949	0.853	0.956
Positive predictive value	0.864	0.840	0.917	0.889	0.808	0.867
Negative predictive value	0.973	0.882	0.936	0.902	0.879	0.977
Balanced accuracy	0.937	0.861	0.882	0.874	0.846	0.942
F-measure	0.905	0.840	0.846	0.842	0.824	0.897

**Table 4.** Confusion matrix and performance measures for SVM classification of wheat genotypes using Sigmoid and Degree-1 polynomial kernels.

were misclassified in medium class, 2 medium yield genotypes in low class, 2 medium yield genotypes in high class and 1 high yield genotype were mislabelled in medium class. The average sensitivity, specificity, PPV, NPV, balanced accuracy and F-measures were calculated and are equal to 85.9% (85.6%), 92.8% (91.9%), 87.4% (85.5%), 93.0% (91.9%), 89.3% (88.7%) and 86.4% (85.4%), respectively for sigmoid (degree-1 polynomial) kernel functions. Sensitivity, negative predictive value, balanced accuracy and F-values were found higher for the low yield class, while specificity and positive predictive values were larger for the high yield class for sigmoid kernel. All the performance measures provide better results in high yield class with degree-1 polynomial, except the positive predictive value which was higher in low yield class.

Table 5 shows the confusion matrix of the detailed classification results of degree-2 and degree-3 polynomial kernels for the test set of yield classes. Almost all the misclassified genotypes fall to the adjacent class with none falling to the distant class in case of degree-3 polynomial kernel but for degree-2 polynomial kernel, the misclassified genotypes also fall to distant classes as well. It is noted that degree-2 and degree-3 polynomial kernels classified the medium yield class with a highest accuracy of 88% and 96%, respectively. Likewise for low and high yield classes, the classification accuracies were comparatively less at 45% & 35.7% and 65% & 64.3% for degree-2 and degree-3 polynomial kernels, respectively. For degree-2 polynomial, the 2 (149th and 191st) genotypes of low class were misclassified to high class, may be due to their higher values for variables DTH, DTA, DTM, GFD, PH, PL, SL and SS as compared to the mean of high class. Likewise, the 8 (17th, 100th, 113th, 119th, 123rd, 167th, 244th and 276th) genotypes of high class were mislabelled to distant low class, as the mean values of some variables for these genotypes were lower than the mean of low yield class. The average sensitivity, specificity, PPV, NPV, balanced accuracy and F-measures were obtained as 56.2% (75.1%), 79.5% (87.5%), 58.1% (85.6%), 81.4% (90.1%), 67.9% (81.3%) and 56.0% (77.5%), respectively for degree-2 (degree-3) polynomial kernels.

The confusion matrix obtained for the EWA approach is shown in Table 6. The performance of the classification model in classifying three yield classes are presented in this matrix. The genotypes in actual classes corresponding to low, medium and high yield are shown in first, second and third columns, respectively. The

Performance statistics	Actual					
	Degree-2 polynomial kernel			Degree-3 polynomial kernel		
	L	M	H	L	M	H
Prediction						
L	9	1	8	13	0	0
M	9	22	1	7	24	5
H	2	2	5	0	1	9
Sensitivity	0.450	0.880	0.357	0.650	0.960	0.643
Specificity	0.769	0.706	0.911	1.000	0.647	0.978
Positive predictive value	0.500	0.688	0.556	1.000	0.667	0.900
Negative predictive value	0.732	0.889	0.820	0.848	0.957	0.898
Balanced accuracy	0.610	0.793	0.634	0.825	0.804	0.810
F-measure	0.474	0.772	0.435	0.788	0.787	0.750

**Table 5.** Confusion matrix and performance measures for SVM classification of wheat genotypes using Degree-2 and Degree-3 polynomial kernels.



Performance statistics	Actual		
	L	M	H
Prediction			
L	19	2	0
M	1	23	0
H	0	0	14
Sensitivity	0.950	0.920	1.000
Specificity	0.949	0.971	1.000
Positive predictive value	0.905	0.958	1.000
Negative predictive value	0.974	0.943	1.000
Balanced accuracy	0.949	0.945	1.000
F-measure	0.927	0.939	1.000

**Table 6.** Confusion matrix and performance measures for EWA approach in wheat genotypes.

genotypes in predicted classes corresponding to low, medium and high yield are shown in first, second and third rows respectively. Nineteen genotypes of actual low yield were predicted as low yield and 1 genotype of low yield was wrongly predicted as medium yield. Twenty-three genotypes of medium yield were predicted correctly in the same medium yield class and 2 genotypes of medium yield class were misclassified in low yield class. The total number of actual high yield genotypes was predicted correctly without any misclassified genotypes. The average sensitivity, specificity, PPV, NPV, balanced accuracy and F-measures were observed as 95.7%, 97.3%, 95.4%, 97.2%, 96.5% and 95.5%, respectively. All performance measures were higher for the high yield class in comparison to other two classes with 100% accuracy.

### Comparative performance of optimization approaches for SVM classification of wheat genotypes

The parameter values which are most important for the improvement of a SVM model are the regularization or penalty parameter ( $C$ ), kernel function and values of kernel parameters. The difficulty encountered in the SVM model is how to select the values of these hyper-parameters. To overcome from this situation, optimization approaches (GS, RS, GA, DE and PSO) for SVM has been compared in this study. The comparison between these five algorithms for optimization of RBF-SVM parameters for classification of wheat genotypes constitute the second phase of the study. For both  $C$  and  $\sigma$  parameters, the search range was  $[10^{-2}$  to  $10^3]$  and maximization of accuracy was used as fitness function in all the approaches. Tournament selection, local arithmetic crossover and uniform random mutation strategies were used as operators in GA. The other parameters for GA were as follows: 1000 number of iterations, population size of 100, crossover rate 0.8, mutation rate 0.1 and elite size of 2. Among the plenty of strategies of DE, local-to-best/1/bin was employed in this study. The other parameters for DE were: 1000 number of iterations, population size ( $N_p$ ) = 100, step size ( $F$ ) = 0.8 and crossover rate ( $CR$ ) = 0.9. Likewise, the parameters for PSO in this study were: number of iterations = 1000 and swarm size = 50.

The results of optimization algorithms for SVM classification model were generated and are given in Table 7 and Fig. S1. As stated in table, classification accuracies over training and testing datasets obtained for the SVM classifier optimized with PSO approach were 94.2% and 94.9%, respectively. These results were better than the

Classifier	Optimum parameter values	Accuracy		Kappa	
		Training	Testing	Training	Testing
SVM	$C = 1$	0.938	0.905	0.932	0.897
	$\sigma = 1$				
GS-SVM	$C = 50$	0.930	0.892	0.847	0.766
	$\sigma = 0.01$				
RS-SVM	$C = 96.51$	0.934	0.899	0.881	0.818
	$\sigma = 0.009$				
GA-SVM	$C = 472.48$	0.938	0.905	0.915	0.871
	$\sigma = 0.12$				
DE-SVM	$C = 476.95$	0.938	0.905	0.915	0.870
	$\sigma = 0.04$				
PSO-SVM	$C = 863.69$	0.942	0.911	0.949	0.922
	$\sigma = 0.01$				

**Table 7.** Comparative performance of various optimization techniques for SVM classification of wheat genotypes.

accuracies attained by the SVM classification with GS, RS, GA and DE optimization algorithms. In fact, classification accuracies over training (testing) datasets were equal to 93.0% (84.7%) for GS-SVM classifier, 93.4% (88.1%) for RS-SVM classifier, and 93.8% (91.5%) for both GA-SVM and DE-SVM classifiers.

The results of GA-SVM were in agreement with the outcomes of Lessmann et al.<sup>91</sup> who also secured 91.86% prediction accuracy with population size of 10 for Wisconsin breast cancer dataset. Liu and Jiao<sup>92</sup> concluded that GA-SVM can access the bridge damage conditions with high accuracy than the RBF kernel. The outcomes of DE-SVM were in line with the results of Bhadra et al.<sup>93</sup> who also achieved the 91% prediction accuracy for Australian dataset and 91.62% testing accuracy for libras movement datasets of UCI machine learning repository.

The best optimization results were obtained when  $C = 50$  and  $\sigma = 0.01$  for GS-SVM,  $C = 96.51$  and  $\sigma = 0.009$  for RS-SVM,  $C = 472.48$  and  $\sigma = 0.12$  for GA-SVM,  $C = 476.95$  and  $\sigma = 0.04$  for DE-SVM, and  $C = 863.69$  and  $\sigma = 0.01$  for PSO-SVM. When we look at these results, we find that there was not a significant improvement in the classification accuracies over training dataset with these optimization algorithms but the accuracy gains of 1.7% was achieved in testing set for PSO-SVM approach. The PSO-SVM technique has set more suitable parameters, provided with the higher classification accuracy in the dataset. The result seems to be in confirmation with what was seen in other areas of application; further proving the superiority of SVM classification based on PSO algorithm compared to other optimization algorithms when working with the high dimension feature space.

Sen et al.<sup>94</sup> also found the classification accuracies of 94% for training set with PSO-SVM model for tumour classification of 22 normal and 40 colon tumour tissues. The results demonstrated that the modified PSO was a useful tool for gene selection and mining high dimension data. Melgani and Bazi<sup>85</sup> optimized the SVM classifier by searching for the best value of the parameters using PSO. The experiment was conducted on the ECG data to classify five kinds of abnormal waveforms and normal beats. The obtained results clearly confirmed the superiority of SVM approach as compared to k-NN and RBF NN. The PSO-SVM yielded an overall accuracy of 89.72% against 82.34%, 83.70% and 85.98% for RBF neural networks, k-NN and SVM classifiers, respectively.

Lin et al.<sup>42</sup> also developed the PSO based approach for parameter determination of the SVM and feature selection. Experimental results demonstrated that the classification accuracies of the developed approach surpass the accuracies of grid search and the PSO-SVM approach had a similar result to GA-SVM. When dealing with fault diagnosis of sensors, Chenglin et al.<sup>95</sup> found a higher average diagnostic accuracy of 93.05% for PSO-SVM with RBF kernel. Bonah et al.<sup>66</sup> concluded that PSO-SVM offers best performance as compared to GA-SVM and GS-SVM for classification of 5 bacterial foodborne pathogens collected from electronic nose dataset.

Tables 8 and 9 depict the confusion matrix and performance measures obtained from the outcomes on test dataset. For the approaches GS-SVM, RS-SVM and DE-SVM, high yield class outperformed other two classes in terms of all the performance measures. In case of GA-SVM, we observed that the performance of high yield class was better than that of the other classes in terms of sensitivity, negative predictive value, balanced accuracy, and F-measure, while medium yield class provided better results in terms of specificity and positive predictive values. Though, for the PSO-SVM, high yield class performed the best over other classes in terms of sensitivity, specificity, negative predictive value, balanced accuracy and F-measure, whereas positive predictive value obtained using medium yield class performed better. Average positive predictive value of 91.14% was generated for GA-SVM while average sensitivity of 92.66% was achieved for DE-SVM. Average positive predictive value of 91.14% was generated for GA-SVM and similar positive predictive value of 92.08% for breast cancer data was found by Liu and Fu<sup>96</sup> with this approach. For DE-SVM, average sensitivity of 92.66% was generated and Bhadra et al.<sup>93</sup> also attained the similar sensitivity value of 92% for an Australian dataset with this approach. Sen et al.<sup>94</sup> obtained the 95% sensitivity in case of PSO-SVM approach, which was in line with the outcomes of this study.

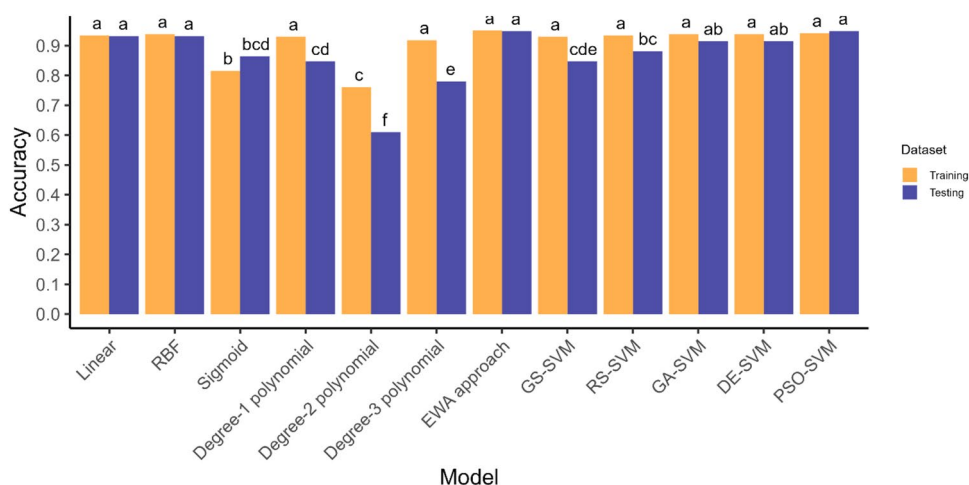
Overall comparison of all the classification models for wheat data are presented in Fig. 6 which shows that the performance of RBF kernel was better than other kernel counterparts in terms of classification accuracies over training as well as testing datasets. Letters a, b, c, d, e and f on bar graph are there to indicate the difference between different models and their accuracies across training and testing datasets at the  $p < 0.05$ . The ensemble approach EWA outperforms others with 95.1% training accuracy followed by PSO-SVM (94.2%). While for

Performance statistics	Actual					
	GS-SVM (RS-SVM)			GA-SVM		
	L	M	H	L	M	H
Prediction						
L	17 (17)	4 (3)	0 (0)	19	2	0
M	3 (3)	20 (21)	1 (0)	1	21	0
H	0 (0)	1 (1)	13 (14)	0	2	14
Sensitivity	0.850 (0.850)	0.800 (0.840)	0.929 (1.000)	0.950	0.840	1.000
Specificity	0.897 (0.923)	0.882 (0.912)	0.978 (0.978)	0.949	0.971	0.956
Positive predictive value	0.810 (0.850)	0.833 (0.875)	0.929 (0.933)	0.905	0.955	0.875
Negative predictive value	0.921 (0.923)	0.857 (0.886)	0.978 (1.000)	0.974	0.892	1.000
Balanced accuracy	0.874 (0.887)	0.841 (0.876)	0.953 (0.989)	0.949	0.905	0.978
F-measure	0.829 (0.850)	0.816 (0.857)	0.929 (0.966)	0.927	0.894	0.933

**Table 8.** Confusion matrix and performance measures for SVM classification of wheat genotypes using GS, RS and GA algorithms.

Performance statistics	Actual					
	DE-SVM			PSO-SVM		
	L	M	H	L	M	H
Prediction						
L	18	2	0	19	1	0
M	2	22	0	1	23	0
H	0	1	14	0	1	14
Sensitivity	0.900	0.880	1.000	0.950	0.920	1.000
Specificity	0.949	0.941	0.978	0.974	0.971	0.978
Positive predictive value	0.900	0.917	0.933	0.950	0.958	0.933
Negative predictive value	0.949	0.914	1.000	0.974	0.943	1.000
Balanced accuracy	0.924	0.911	0.989	0.962	0.945	0.989
F-measure	0.900	0.898	0.966	0.950	0.939	0.966

**Table 9.** Confusion matrix and performance measures for SVM classification of wheat genotypes using DE and PSO algorithms.



**Fig. 6.** Overall comparison of various kernels, ensemble approach and optimization algorithms for classification of wheat genotypes using SVM. Letters a, b, c, d, e and f on bar graph are there to indicate the difference between different models and their accuracies across training and testing datasets at the  $p < 0.05$ .

testing data set, the EWA approach and PSO-SVM performed well with 94.9% accuracy. The lowest training (76.1%) and testing (61.0%) accuracies were obtained for Degree-2 polynomial kernel function.

The results of one-way ANOVA for training as well as testing datasets depicted that the null hypothesis gets rejected with a very high significance level (Supplementary material Table S2). So, the difference in the accuracies of all twelve classifiers was found statistically significant. The Supplementary material Table S3 illustrated the application of Tukey HSD test on this result. These results show that the accuracies of all the classifiers were found statistically significant with the accuracies of Sigmoid and Degree-2 polynomial kernels in case of training dataset. But, in case of testing dataset, the mean comparisons of classifiers EWA and PSO-SVM were found statistically significant with the accuracies of Sigmoid, Degree-1 polynomial, Degree-2 polynomial, Degree-3 polynomial, GS and RS classifiers.

## Conclusions and future research

Nowadays, SVM has attracted much attention as an effective problem-solving technique for real-world classification tasks. However, the performance of SVM strongly depends on the selection of the corresponding kernel function and the value of the associated parameters. Based on the results, we strongly recommend using SVM classification with RBF kernels for wheat genotypes due to their better generalization ability than linear, sigmoid and polynomial kernels. This ability usually gives them high classification accuracy. The main emphasis of this research lies in the ensemble approach EWA and the optimization algorithms that aim to combine the results of different base classifiers and optimize the accuracy of SVM classifiers, respectively. To achieve this, the study also compares optimization approaches.

The performance of the RBF kernel was better than other kernel features in terms of classification accuracy on both training and test datasets. Ensembling with the weighted accuracy approach outperformed the results of

the individual kernel functions. For classification accuracy in the test data set, the PSO-SVM approach produced a gain of 1.7%, compared to the RBF kernel results. The results of this study were obtained with the RBF kernel because its accuracy was the highest among all kernels tested in this study.

However, the parameters of other kernel functions can also be optimized using these approaches and/or other approaches can be compared. Some other classification techniques, such as Fisher's linear discriminant function, k-nearest neighbor classifier, Bayesian networks, decision trees and artificial neural networks, among others, can also be compared with optimization-based SVM classification in the future. The recent advancements in optimization algorithms, such as the Grey Wolf Optimizer (GWO) and the Salp Swarm Algorithm (SSA), can be utilized for comparative analysis in future studies. The results of this study can be generalized by using this type of genotypic data for other crops.

## Data availability

The data supporting this study's findings are available from the first author or Dinesh Kumar Vishwakarma upon request.

Received: 28 May 2024; Accepted: 3 September 2024

Published online: 30 September 2024

## References

- Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. & Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **408**, 189–215 (2020).
- Zhang, Y., Cao, G., Wang, B. & Li, X. A novel ensemble method for k-nearest neighbor. *Pattern Recognit.* **85**, 13–25 (2019).
- Marcot, B. G. & Penman, T. D. Advances in Bayesian network modelling: Integration of modelling technologies. *Environ. Model. Softw.* **111**, 386–393 (2019).
- Huang, D.-S., Ip, H. H. S., Law, K. C. K. & Chi, Z. Zeroing polynomials using modified constrained neural network approach. *IEEE Trans. Neural Netw.* **16**, 721–732 (2005).
- Trabelsi, A., Elouedi, Z. & Lefevre, E. Decision tree classifiers for evidential attribute values and class labels. *Fuzzy Sets Syst.* **366**, 46–62 (2019).
- Huang, D.-S. & Du, J.-X. A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. *IEEE Trans. Neural Netw.* **19**, 2099–2115 (2008).
- Zhang, M., Qu, H., Xie, X. & Kurths, J. Supervised learning in spiking neural networks with noise-threshold. *Neurocomputing* **219**, 333–349 (2017).
- Vapnik, V. *Statistical Learning Theory* Vol. 1 (Wiley, 1998).
- Vapnik, V. N. *The Nature of Statistical Learning Theory* 286 (Springer, 1995).
- Liang, X., Zhu, L. & Huang, D.-S. Multi-task ranking SVM for image cosegmentation. *Neurocomputing* **247**, 126–136 (2017).
- Zhu, F., Chen, W., Guo, F. & Zhang, X. Combining context connectivity and behavior association to develop an indoor/outdoor context detection model with smartphone multisensor fusion. *IEEE Internet Things J.* **11**, 2883–2898 (2024).
- Hu, J., Peng, Y., Lin, Q., Liu, H. & Zhou, Q. An ensemble weighted average conservative multi-fidelity surrogate modeling method for engineering optimization. *Eng. Comput.* **38**, 2221–2244 (2022).
- Suh, M.-S. *et al.* Development of new ensemble methods based on the performance skills of regional climate models over South Korea. *J. Clim.* **25**, 7067–7082 (2012).
- Ghosh, P. *et al.* Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access* **9**, 19304–19326 (2021).
- Ardjani, F., Sadouni, K. & Benyettou, M. Optimization of SVM MultiClass by particle swarm (PSO-SVM). In *2010 2nd International Workshop on Database Technology and Applications* 1–4 (IEEE, 2010). <https://doi.org/10.1109/DBTA.2010.5658994>.
- Dudzic, W., Kawulok, M. & Nalepa, J. Evolutionarily-tuned support vector machines. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* 165–166 (ACM, 2019). <https://doi.org/10.1145/3319619.3321924>.
- Huang, C.-L. & Dun, J.-F. A distributed PSO-SVM hybrid system with feature selection and parameter optimization. *Appl. Soft Comput.* **8**, 1381–1391 (2008).
- Dudzic, W., Nalepa, J. & Kawulok, M. Evolving data-adaptive support vector machines for binary classification. *Knowl.-Based Syst.* **227**, 107221 (2021).
- Hitam, N. A., Ismail, A. R. & Saeed, F. An optimized support vector machine (SVM) based on particle swarm optimization (PSO) for cryptocurrency forecasting. *Procedia Comput. Sci.* **163**, 427–433 (2019).
- Luo, Z., Xiang, M. & Zhang, X. Multi-class wavelet SVM classifiers using quantum particles swarm optimization algorithm. In *2008 International Symposium on Computational Intelligence and Design* 278–281 (IEEE, 2008). <https://doi.org/10.1109/ISCID.2008.93>.
- Liao, R. J., Zheng, H. B., Grzybowski, S. & Yang, L. J. A multiclass SVM-based classifier for transformer fault diagnosis using a particle swarm optimizer with time-varying acceleration coefficients. *Int. Trans. Electr. Energy Syst.* **23**, 181–190 (2013).
- Li, J. & Li, B. Parameters selection for support vector machine based on particle swarm optimization. In *International Conference on Intelligent Computing* (eds. Huang, D., Bevilacqua, V. & Premaratne, P.) 41–47 (Intelligent Computing Theory. ICIC 2014. Lecture Notes in Computer Science, 2014). [https://doi.org/10.1007/978-3-319-09333-8\\_5](https://doi.org/10.1007/978-3-319-09333-8_5).
- Nalepa, J., Dudzic, W. & Kawulok, M. Memetic evolution of training sets with adaptive radial basis kernels for support vector machines. In *2020 25th International Conference on Pattern Recognition (ICPR)* 5503–5510 (IEEE, 2021). <https://doi.org/10.1109/ICPR48806.2021.9412495>.
- Nugraha, Y. R., Wibawa, A. P. & Zaeni, I. A. E. Particle swarm optimization—support vector machine (PSO-SVM) algorithm for journal rank classification. In *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)* 69–73 (IEEE, 2019). <https://doi.org/10.1109/IC2IE47452.2019.8940822>.
- Sheela, M. S. & Arun, C. A. Hybrid PSO-SVM algorithm for Covid-19 screening and quantification. *Int. J. Inf. Technol.* **14**, 2049–2056 (2022).
- Golcuk, A. & Yasar, A. Classification of bread wheat genotypes by machine learning algorithms. *J. Food Compos. Anal.* **119**, 105253 (2023).
- Olgun, M. *et al.* Wheat grain classification by using dense SIFT features with SVM classifier. *Comput. Electron. Agric.* **122**, 185–190 (2016).
- Guevara-Hernandez, F. & Gomez-Gil, J. A machine vision system for classification of wheat and barley grain kernels. *Span. J. Agric. Res.* **9**, 672 (2011).
- Gülmezoğlu, M. B. & Gülmezoğlu, N. Classification of bread wheat varieties and their yield characters with the common vector approach. In *International Conference on Chemical, Environmental and Biological Sciences (CEBS-2015) March 18–19, 2015 Dubai*



- (UAE) 120–123 (International Institute of Chemical, Biological & Environmental Engineering, 2015). <https://doi.org/10.15242/IICBE.C0315090>.
30. Majumdar, S. & Jayas, D. S. Classification of cereal grains using machine vision: IV. Combined morphology, color, and texture models. *Trans. ASAE* **43**, 1689–1694 (2000).
  31. Majumdar, S. & Jayas, D. S. Classification of cereal grains using machine vision: II. Colormodels. *Trans. ASAE* **43**, 1677–1680 (2000).
  32. Majumdar, S. & Jayas, D. S. Classification of cereal grains using machine vision: I. Morphology models. *Trans. ASAE* **43**, 1669–1675 (2000).
  33. Yasar, A., Golcuk, A. & Sari, O. F. Classification of bread wheat varieties with a combination of deep learning approach. *Eur. Food Res. Technol.* **250**, 181–189 (2024).
  34. Kılıçarslan, S. & Kılıçarslan, S. A comparative study of bread wheat varieties identification on feature extraction, feature selection and machine learning algorithms. *Eur. Food Res. Technol.* **250**, 135–149 (2024).
  35. Ismail, A. *et al.* A novel deep learning-based model for classification of wheat gene expression. *Comput. Syst. Sci. Eng.* **48**, 273–285 (2024).
  36. Zhang, J. & Yang, H. Bounded quantile loss for robust support vector machines-based classification and regression. *Expert Syst. Appl.* **242**, 122759 (2024).
  37. Yeganeh, A., Abbasi, S. A., Shongwe, S. C., Malela-Majika, J.-C. & Shadman, A. R. Evolutionary support vector regression for monitoring Poisson profiles. *Soft Comput.* <https://doi.org/10.1007/s00500-023-09047-2> (2023).
  38. Chauhan, V. K., Dahiya, K. & Sharma, A. Problem formulations and solvers in linear SVM: A review. *Artif. Intell. Rev.* **52**, 803–855 (2019).
  39. Singh, V. K. *et al.* Novel genetic algorithm (GA) based hybrid machine learning-pedotransfer function (ML-PTF) for prediction of spatial pattern of saturated hydraulic conductivity. *Eng. Appl. Comput. Fluid Mech.* **16**, 1082–1099 (2022).
  40. Yalsavar, M. *et al.* Kernel parameter optimization for support vector machine based on sliding mode control. *IEEE Access* **10**, 17003–17017 (2022).
  41. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
  42. Lin, S.-W., Ying, K.-C., Chen, S.-C. & Lee, Z.-J. Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Syst. Appl.* **35**, 1817–1824 (2008).
  43. Araujo, M. & New, M. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42–47 (2007).
  44. Cherkassky, V. & Ma, Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Netw.* **17**, 113–126 (2004).
  45. Kayhomayoon, Z., Babsaeian, F., Ghordoyee Milan, S., Arya Azar, N. & Berndtsson, R. A combination of metaheuristic optimization algorithms and machine learning methods improves the prediction of groundwater level. *Water* **14**, 751 (2022).
  46. Seifi, A., Ehteram, M., Singh, V. P. & Mosavi, A. Modeling and uncertainty analysis of groundwater level using six evolutionary optimization algorithms hybridized with ANFIS, SVM, and ANN. *Sustainability* **12**, 4023 (2020).
  47. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, 1999).
  48. Rifkin, R. & Klautau, A. In defense of one-vs-all classification. *J. Mach. Learn. Res.* **5**, 101–141 (2004).
  49. Hastie, T. & Tibshirani, R. Classification by pairwise coupling. In *Advances in Neural Information Processing Systems* (eds. Jordan, M. I., Kearns, M. J. & Solla, S. A.) vol. 10 507–513 (1997).
  50. Pal, M. Multiclass approaches for support vector machine based land cover classification. *Neural Evol. Comput.* <https://doi.org/10.48550/arXiv.0802.2411> (2008).
  51. Wang, H., Zheng, B., Yoon, S. W. & Ko, H. S. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Oper. Res.* **267**, 687–699 (2018).
  52. Ren, Y., Zhang, L. & Suganthan, P. N. Ensemble classification and regression—recent developments, applications and future directions. *IEEE Comput. Intell. Mag.* **11**, 41–53 (2016).
  53. Rossi, A. L. D. & Carvalho, A. C. P. L. F. de. Bio-inspired Optimization Techniques for SVM Parameter Tuning. In *2008 10th Brazilian Symposium on Neural Networks* 57–62 (IEEE, 2008). <https://doi.org/10.1109/SBRN.2008.28>.
  54. Cho, M.-Y. & Hoang, T. T. Feature selection and parameters optimization of SVM using particle swarm optimization for fault classification in power distribution systems. *Comput. Intell. Neurosci.* **2017**, 1–9 (2017).
  55. Mantovani, R. G., Rossi, A. L. D., Vanschoren, J., Bischl, B. & de Carvalho, A. C. P. L. F. Effectiveness of Random Search in SVM hyper-parameter tuning. In *2015 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2015). <https://doi.org/10.1109/IJCNN.2015.7280664>.
  56. Wu, X., Pan, J. & Zhu, X. Optimizing the ecological source area identification method and building ecological corridor using a genetic algorithm: A case study in Weihe River Basin, NW China. *Ecol. Inform.* **80**, 102519 (2024).
  57. Fernandez, M., Caballero, J., Fernandez, L. & Sarai, A. Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). *Mol. Divers.* **15**, 269–289 (2011).
  58. Kumar, A. *et al.* Development of novel hybrid models for prediction of drought- and stress-tolerance indices in teosinte introgressed maize lines using artificial intelligence techniques. *Sustainability* **14**, 2287 (2022).
  59. Mallipeddi, R., Suganthan, P. N., Pan, Q. K. & Tasgetiren, M. F. Differential evolution algorithm with ensemble of parameters and mutation strategies. *Appl. Soft Comput.* **11**, 1679–1696 (2011).
  60. Li, J., Ding, L. & Li, B. Differential evolution-based parameters optimisation and feature selection for support vector machine. *Int. J. Comput. Sci. Eng.* **13**, 355 (2016).
  61. Vesterstrom, J. & Thomsen, R. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)* vol. 2 1980–1987 (IEEE, 2004).
  62. Kennedy, J. & Eberhart, R. Particle swarm optimization. In *Proceedings of ICNN'9—International Conference on Neural Networks* vol. 4 1942–1948 (IEEE, 1995).
  63. Al-Thanoon, N. A., Qasim, O. S. & Algarni, Z. Y. A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics. *Chemom. Intell. Lab. Syst.* **184**, 142–152 (2019).
  64. Li, X., Wu, S., Li, X., Yuan, H. & Zhao, D. Particle swarm optimization-support vector machine model for machinery fault diagnoses in high-voltage circuit breakers. *Chinese J. Mech. Eng.* **33**, 6 (2020).
  65. Subasi, A. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. *Comput. Biol. Med.* **43**, 576–586 (2013).
  66. Bonah, E. *et al.* Electronic nose classification and differentiation of bacterial foodborne pathogens based on support vector machine optimized with particle swarm optimization algorithm. *J. Food Process Eng.* **42**, e13236 (2019).
  67. Dubey, B. P., Bhagwat, S. G., Shouche, S. P. & Sainis, J. K. Potential of artificial neural networks in varietal identification using morphometry of wheat grains. *Biosyst. Eng.* **95**, 61–67 (2006).
  68. Masoumiasl, A., Amiri-Fahliani, R. & Khoshroo, A. R. Some local and commercial rice (*Oryza sativa* L.) varieties comparison for aroma and other qualitative properties. *Int. J. Agric. Crop Sci.* **5**, 2184–2189 (2013).

69. Zhang, H., Paliwal, P., Jayas, D. S. & White, N. D. G. Classification of fungal infected wheat kernels using near-infrared reflectance hyperspectral imaging and support vector machine. *Trans. ASABE* **50**, 1779–1785 (2007).
70. Yao, Q. *et al.* Application of support vector machine for detecting rice diseases using shape and color texture features. In *2009 International Conference on Engineering Computation* 79–83 (IEEE, 2009). <https://doi.org/10.1109/ICEC.2009.73>.
71. Jian, Z. & Wei, Z. Support vector machine for recognition of cucumber leaf diseases. In *2010 2nd International Conference on Advanced Computer Control* vol. 5 264–266 (IEEE, 2010).
72. Dubey, S. R. & Jalal, A. S. Detection and classification of apple fruit diseases using complete local binary patterns. In *2012 Third International Conference on Computer and Communication Technology* 346–351 (IEEE, 2012). <https://doi.org/10.1109/ICCCT.2012.76>.
73. Sengupta, S. & Lee, W. S. Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. *Biosyst. Eng.* **117**, 51–61 (2014).
74. Bhange, M. & Hingoliwala, H. A. Smart farming: Pomegranate disease detection using image processing. *Procedia Comput. Sci.* **58**, 280–288 (2015).
75. Chung, C.-L. *et al.* Detecting Bakanae disease in rice seedlings by machine vision. *Comput. Electron. Agric.* **121**, 404–411 (2016).
76. Padol, P. B. & Yadav, A. A. SVM classifier based grape leaf disease detection. In *2016 Conference on Advances in Signal Processing (CASP)* 175–179 (IEEE, 2016). <https://doi.org/10.1109/CASP.2016.7746160>.
77. Sharma, D. *et al.* Mapping quantitative trait loci associated with grain filling duration and grain number under terminal heat stress in bread wheat (*Triticum aestivum* L.). *Plant Breed.* **135**, 538–545 (2016).
78. Jenks, G. F. Generalization in statistical mapping. *Ann. Assoc. Am. Geogr.* **53**, 15–26 (1963).
79. Japkowicz, N. & Shah, M. *Evaluating Learning Algorithms: A Classification Perspective* (Cambridge University Press, 2011).
80. Powers, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Mach. Learn.* <https://doi.org/10.48550/arXiv.2010.16061> (2020).
81. Sokolova, M., Japkowicz, N. & Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science* (eds Sattar, A. & Kang, B.) 1015–1021 (Springer, 2006). [https://doi.org/10.1007/11941439\\_114](https://doi.org/10.1007/11941439_114).
82. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
83. Zhang, J., Wang, Y., Sun, Y. & Li, G. Strength of ensemble learning in multiclass classification of rockburst intensity. *Int. J. Numer. Anal. Methods Geomech.* **44**, 1833–1853 (2020).
84. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **17**, 168–192 (2021).
85. Melgani, F. & Bazi, Y. Classification of electrocardiogram signals with support vector machines and particle swarm optimization. *IEEE Trans. Inf. Technol. Biomed.* **12**, 667–677 (2008).
86. Zhang, Y. & Wu, L. Classification of fruits using computer vision and a multiclass support vector machine. *Sensors* **12**, 12489–12505 (2012).
87. Bulanon, D. M., Burks, T. F. & Alchanatis, V. Study on Fruit Visibility for Robotic Harvesting. In *2007 Minneapolis, Minnesota, June 17–20, 2007* (American Society of Agricultural and Biological Engineers, 2007). <https://doi.org/10.13031/2013.23428>.
88. Lu, Q., Cai, J., Zhao, J., Wang, F. & Tang, M. Real-time recognition of citrus on trees in natural scene. *Nongye Jixie Xuebao = Trans. Chin. Soc. Agric. Mach.* **41**, 170–185 (2010).
89. Zhang, W., Yoshida, T. & Tang, X. Text classification based on multi-word with support vector machine. *Knowl.-Based Syst.* **21**, 879–886 (2008).
90. Manurung, J., Mawengkang, H. & Zamzami, E. Optimizing support vector machine parameters with genetic algorithm for credit risk assessment. *J. Phys. Conf. Ser.* **930**, 012026 (2017).
91. Lessmann, S., Stahlbock, R. & Crone, S. F. Genetic algorithms for support vector machine model selection. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings* 3063–3069 (IEEE, 2006). <https://doi.org/10.1109/IJCNN.2006.247266>.
92. Liu, H.-B. & Jiao, Y.-B. Application of genetic algorithm-support vector machine (GA-SVM) for damage identification of bridge. *Int. J. Comput. Intell. Appl.* **10**, 383–397 (2011).
93. Bhadra, T., Bandyopadhyay, S. & Maulik, U. Differential evolution based optimization of SVM parameters for meta classifier design. *Procedia Technol.* **4**, 50–57 (2012).
94. Shen, Q., Shi, W.-M., Kong, W. & Ye, B.-X. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. *Talanta* **71**, 1679–1683 (2007).
95. Chenglin, Z., Xuebin, S., Songlin, S. & Ting, J. Fault diagnosis of sensor by chaos particle swarm optimization algorithm and support vector machine. *Expert Syst. Appl.* **38**, 9908–9912 (2011).
96. Liu, X. & Fu, H. PSO-based support vector machine with cuckoo search technique for clinical disease diagnoses. *Sci. World J.* **2014**, 1–7 (2014).

## Acknowledgements

The authors extend their appreciation to the Deanship of Research and Graduate studies at King Khalid University for funding this work through large group Research Project under grant number RGP.2/67/45. Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R584), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

## Author contributions

Conceptualization and Methodology: Mujahid Khan, B. K. Hooda and Arpit Gaur; Data curation, Formal analysis and Investigation: Mujahid Khan, Arpit Gaur and Sonia Sheoran; Project administration: Mujahid Khan and B. K. Hooda; Resources: Mujahid Khan, Arpit Gaur and Vikram Singh; Software: Mujahid Khan and B. K. Hooda; Supervision, Validation and Visualization: Mujahid Khan, Arpit Gaur and Dinesh Kumar Vishwakarma; Writing—original draft: Mujahid Khan, B. K. Hooda, Arpit Gaur, Vikram Singh, Yogesh Jindal, Hemender Tanwar, and Sushma Sharma; Writing—review & editing: Mujahid Khan, Arpit Gaur, Dinesh Kumar Vishwakarma, Mohammad Khalid, Ghadah Shukri Albakri, Maha Awjan Alreshidi, Jeong Ryeol Choi and Krishna Kumar Yadav. All authors agree and approve the final version of the manuscript.

## Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.: NRF-2021R1F1A1062849).

## Competing interests

The authors declare no competing interests.

### Consent to participate

Before submitting the paper, all the authors have given consent to publish.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-72056-0>.

**Correspondence** and requests for materials should be addressed to D.K.V. or J.R.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2024