Vol 3 (3) (2022)

Measurement, Control, and Automation

Website: https:// mca-journal.org

ISSN 1859-0551

An accurate machine learning approach for seed germination prediction

Nguyễn Đình Văn[1]*, Đào Trung Kiên[1], Nguyễn Việt Tùng[1], Phạm Thị Ngọc Yến[2]

[1]*MICA Institute & Departement of Communication Engineering, SEEE, HUST*
[2]*MICA Institute & Departement of Automation, SEEE, HUST*
[*]*Corresponding author E-mail: van.nguyendinh@hust.edu.vn*

Abstract

To determine the quality of seeds, researchers often must manually check for seeds germination. The process is cumbersome, time consuming and error-prone since it requires the researcher to manually examine at least a few hundred to thousands of seeds. Hence, an automatic seeds germination prediction solution is required. Over the years, with the help of deep learning methods, some studies have accurately predicted the performance of seeds given just a picture of them. However, one downside of the deep learning approaches is the result does not give more insight into which factors of the seeds' image contribute to a successful germination process. In this paper, we propose a classical machine learning method with a carefully designed features engineering process to both accurately predict seeds germination and give more insight into the relevant factors for a seed's germination process. At 95% prediction precision, the proposed method suggests that relevant factors are: seed's size, the circularity, brightness distribution and its skewness and kurtosis.

Keywords*: Feature engineering; Machine learning; Seeds germination prediction; Smart agriculture; Seeds quality;*

Abbreviations

ANN	Artificial Neural Network
ITSA	International Seed Testing Association
RF	Random Forest
RGB	Red Green Blue
SVC	Support Vector Machine
SVIS	Seed Vigor Imaging System

Tóm tắt

Để xác định chất lượng của hạt giống, các nhà nghiên cứu thường phải kiểm tra khả năng nảy mầm của hạt một cách thủ công. Quá trình này rất rườm rà, tốn thời gian do dễ xảy ra lỗi vì cần thực hiện kiểm tra ít nhất từ vài trăm đến hàng nghìn hạt giống bằng mắt thường. Do đó, một giải pháp dự đoán hạt nảy mầm tự động là cần thiết. Trong những năm qua, với sự trợ giúp của các phương pháp học sâu, một số nghiên cứu đã dự đoán chính xác hiệu suất của hạt giống chỉ với một hình ảnh về chúng. Tuy nhiên, một mặt trái của các phương pháp học sâu là kết quả không cung cấp thêm thông tin chi tiết về yếu tố nào trong hình ảnh của hạt giống góp phần vào quá trình nảy mầm thành công. Trong bài báo này, chúng tôi đề xuất một phương pháp học máy cổ điển với quy trình kỹ thuật tính năng được thiết kế cẩn thận để dự đoán chính xác sự nảy mầm của hạt và cung cấp cái nhìn sâu sắc hơn về các yếu tố liên quan cho quá trình nảy mầm của hạt. Với độ chính xác dự báo khả năng nảy mầm khoảng 95%, phương pháp đề xuất gợi ý rằng các yếu tố liên quan đến khả năng nảy mầm của hạt bao gồm: kích thước, độ tròn, phân bố độ sáng, độ lệch và độ nhọn của hạt giống.

1. Introduction

Germination is a crucial attribute of seed quality assessment. It directly impacts the produce yield and quality of the plant. Often, companies must assess seeds germination to meet certain germination standards before distributing to the customers. The process of assessing germination has two parts:

- Seed germinability (the ability to germinate)
- Seed usability (after germination, the seed is usable)

However, germination of a specific seed type varies due to the condition it was produced, harvested, stored, and germinated. There are multiple factors during these phrases that could lead to seed unable to germinate [1]. Therefore, companies must continuously assess seed germination for multiple seed lots to ensure quality of the product. Thus, it is critical to develop an automatic seed germination solution. In this paper, we focus on seed germinability prediction process and the relevant factors that contribute to seed germinability using RGB images.

With advancements made in computer visions, numerous attempts have been made to assess seed (and grain) quality by developing non-destructive, automated predicting models that are capable of judging each specific seed rather than just a statistical result [2]-[8]. At the cores of these studies, classical machine learning or deep learning approaches enable high accuracy prediction of seed germination. While deep learning approach gained a lot more attention recently due to its precision, the downside of it is clear. A deep learning model can predict up to 98% accuracy seed germinability given enough samples but does not provide insight on which

Received: 21 November 2022; Accepted: 22 December 2022.

factors are relevant to the germinability of a seed. This is because the way deep learning utilizes neural network algorithms to automatically extract features out of the RGB image of a seed [9]. In contrast, classical machine learning approach needs a feature engineering process where features are extracted from RGB images of seed. The process is difficult since there is no guarantee of a successful machine learning model given a set of features amongst an infinite combination of possible features [10]. Still, having a defined set of relevant features would offer much more insight into the germination process of a particular seed.

Given the physical appearance and chemical characteristics of a seed can be extracted from RGB images [11]-[14], it is feasible to extract a set of relevant features from RGB images of seeds to predict the germinability. In this study, we propose a set of features extracted from different types of seeds and a machine learning model to accurately predict seeds germination.

The paper has 5 sections organized as follows: an introduction about the problem in section 1; the state of the art of the matter is presented in section 2. Section 3 provides details about methodology to carry out the study. Section 4 presents results and evaluation of models proposed in the study. A conclusion can be seen in section 5.

2. State of the art

2.1. Conventional methods

The conventional method for testing seed quality often includes seed vigor tests. These tests can potentially show all properties for a seed which determines seed lots germinability in a wide range of environments. Most of these tests are developed by the International Seed Testing Association (ITSA) [15]. However, these tests need to be evaluated manually using different complex standardized procedures for different seeds. Thus, they are not commonly used since it requires wide range of tests, with time intensive protocols [16].

2.2. Computer vision approaches

To automate the process of testing seeds germinability as well as reducing human errors, image processing and analysis techniques are commonly used. Approaches in [2]-[8] shows multiple attempts to use classical image analysis to determine the correlation between features extracted from RGB image of seeds with its germination.

Germinator, for instance, is a software that measure differences in time of seed images to look for indication of germination [2].

Seed Vigor Imaging System (SVIS) [3] processes RGB pixel values of scanned images to calculate the length of seeds. The system making use of flatbed scanner to reduce the illumination or partial occlusion issues met in camera capturing methods.

In some cases, X-ray images or high-resolution spectral detectors are also used to capture different views of seeds. These techniques, however, require costly equipment as well as laboratory conditions to execute [17]-[19].

Still, some of existing industrial solutions which offer most accurate characteristics of seeds often work in destructive methods. This means seeds will be destroyed or partially damaged during evaluation and testing phases [20],[21]. Hence, industrial companies still very much rely on conventional methods when it comes to germination evaluation.

3. Methodology

In our study, we focus on a non-destructive computer vision approach that can be easily adopted by a standardized setting of a farm factory. Not only trying to accurately predict seed germinability, but we also attempt to find relevant features which are important to the germination process. Seed usability was not in the scope of the study and subject to further research.

The method to carry out the study follows closely a standard classification procedure for machine learning in computer vision. There are three major steps:

- Data collection
- Data pre-processing and analysis
- Training and evaluation of machine learning algorithms

3.1. Data collection

Data collection is the most time-consuming, labouring and accuracy intensive task in a machine learning problem. There are 2 aspects of data collection phase:

- The collected dataset's size
- The quality of the dataset: labelling accuracy, noise ratio, etc.

Often, it requires thousands of datapoints for a simple machine learning problem and a degree of tens of thousands is recommended for a deep learning problem. However, all these datapoints must be accurately labelled with minimal random noises (outsides factors that can contribute to the incorrect datapoint). And often, the labelling process is done manually (supervised learning) to prevent any error.

In our study, we focus on two seed types which are popular in the northern of Vietnam which are: a variation of spinach (cải bó xôi) and water spinach (rau muống). The two seed types are shown in Fig.1. Spinach is a long harvesting cycle seed of about 40-45 days for a cycle while water spinach is a medium cycle length with 20-25 days.

(a) (b)

Figure 1: Two seed types chosen for the study: (a): spinach variation (cải bó xôi), (b) water spinach (rau muống).

The process of collecting data for the study is illustrated in Fig.2. There are three phrases of data collection process: image capturing; seed batch processing and analysis; germination outcome checkup & matching.

In image capturing stage, a batch of seeds (often 50 – 60 seeds per batch) is laid on a special color paper with an id number and batch id. Batch sample images can be seen in Fig.3. and Fig. 4. In Fig.3, a spinach batch image is shown with 50 seeds and a blue background. The blue background is designed to be easily removed and facilitate seed segmentation in the image processing stage. Fig.4 shows a batch of water spinach of 60 seeds and removed background.

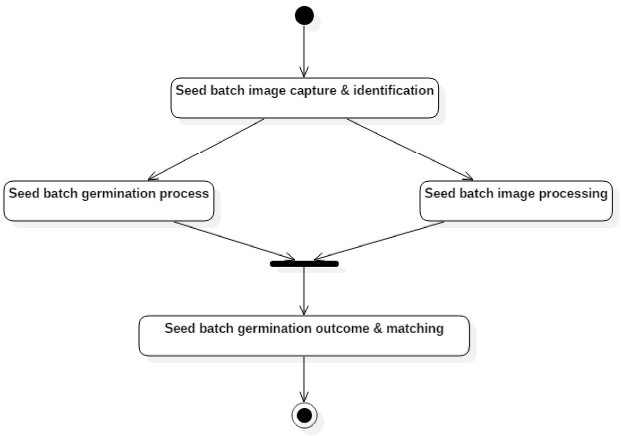

Figure 2: Illustration of data collection process. There are 3 stages: image capturing; seed batch processing and analysis; germination outcome check-up & matching.

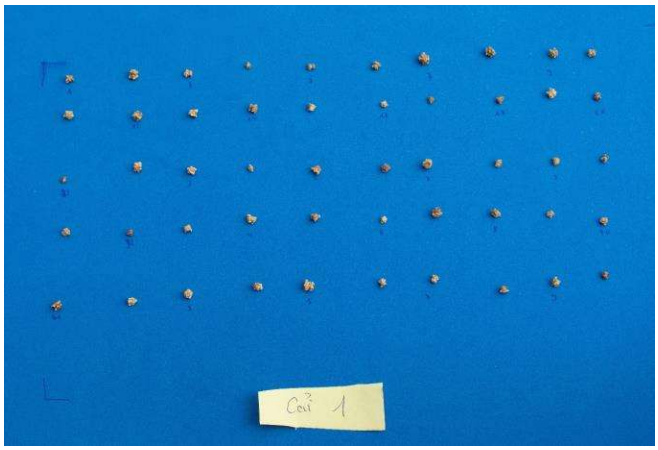

Figure 3: An image of a spinach batch (cải bó xôi) with 50 seeds and blue background.

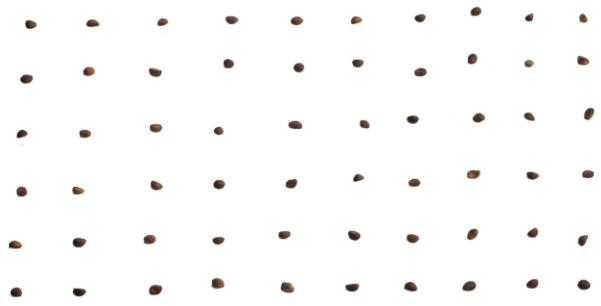

Figure 4: An image of a water spinach batch (rau muống) with 60 seeds and removed background (after background removal).

It is then captured using a high-resolution camera with specification shown in Table 1. The setup is fixed for the entire experiment to ensure all seeds are captured in the same conditions (lighting, background, position, distance to the camera and distortion).

An image of the setup can be seen in Fig. 5 (note that light coming from windows is allowed only to capture the setup. The setup using flashlight to maintain constant lighting environment.) Flashes were not shown in the setup image as it is placed far away to avoid hard lighting and reflection.

Table 1: Specification of equipment and setup for seed batch image capture.

ID	Name	Specification
1	Nikon DSLR D610	CMOS sensor 35.9mm 24.3 Megapixels TTL exposure metering using 2,016-pixel RGB sensor
2	Lens Nikkor 85mm 1.8	Focal lens: 85mm Maximum aperture: f/1.8 Minimum aperture: f/16 Focus distance: 80cm
3	Nikon Speedlight SB-700	Effective flash output distance range: 0.6m – 20m Guide number: 28/92 (ISO 100, m/ft.), 39/128 (ISO 200, m/ft.)
4	Use flash	Yes
5	Focal length	85mm
6	Aperture	f/11
7	Speed	1/160
8	Distance batch seed to camera	90cm

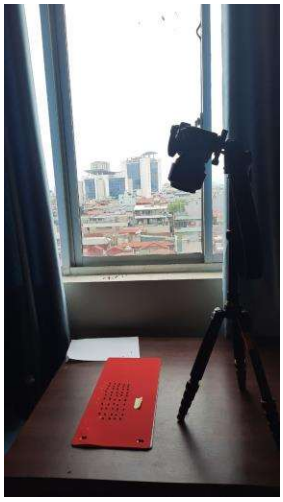

Figure 5: Setup for batch image capturing with camera, tripod and a batch of seed placed on a special paper to absorb hard lighting and facilitate background removal, seed segmentation later.

In total, 100 batches of 50 seeds for spinach were captured and 100 batches of 60 seeds for water spinach were captured. It resulted in 5000 seeds for spinach and 6000 seeds for water spinach.

In the second stage seed batch processing and analysis, seed images will be processed, saved and at the same time, seeds go through the germination process using recommended setup. Each seed will be carefully placed in a numbered tray so it can be identified later. The tray of seeds can be seen in Fig. 6.

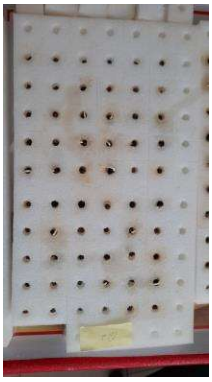

Figure 6: Tray for batch of seeds number 9. The starting seed id is marked with missing piece in bottom right of the image.

The image processing process will be discussed in detail in section 3.2 Data pre-processing and analysis.
In the last stage: germination outcome checkup & matching, after going through the germination process, seeds will be marked with either successful germinated (1) and unsuccessful germinated (0) as the final label for the seed. Since seed ID and seed batch number were carefully kept during the process, it is possible to match the germination outcome and the image of each seed.
The outputs of the data collection process are one dataset of 6000 seeds for water spinach with germination outcome labels and another dataset of 5000 seeds for spinach also with germination outcome labelling.

3.2. Data pre-processing and analysis

The process of data pre-processing and analysis consists of two parts: (1) data cleansing including noise reduction, background removal, seeds segmentation and seeds boundary identification; (2) features selection and extraction.
Images of batch of seeds are fed into a data cleansing module which performs noise reduction, background removal, seeds segmentation and seeds boundary identification. First, noise of images will be reduced using fastNlMeansDenoising algorithm [22]. Then a background removal is done by selecting the pre-defined background colour. An image of transparent background batch of seeds facilitates the seeds segmentation and boundary identification. The sample of seeds boundary can be seen in Fig. 7.

Figure 7: Seeds boundary and center point.

Given the seeds boundary and center point identified, the identification of seeds can be done by sorting seeds using x and then y coordinate. Together with the batch id, every seed is identified and matched with its germination outcome later.

In features selection and extraction task, after reviewing multiple state of the art studies in [2]-[10], selected features are extracted automatically from images of seeds. The selected features are explained in Table 2.

Table 2: Selected features of seed image

ID	Name	Specification
1	Seed area	Calculated in pixels, demonstrate a relative size of seed compare to others of the same type
2	Mean gray value	Calculated based on grayscale image, demonstrate the distribution of brightness in the seed image.
3	Perimeter	Show the perimeter of seed
4	Circularity	Calculate the circularity of a seed with value from 0: non-uniform to 1: perfect circularity
5	Skew	The skewness of the shape of the seed, the third order moment of the mean
6	Kurt	Kurtosis of the seed shape, the fourth order moment of the mean.
7	Aspect Ratio	The ratio of major axis / minor axis in case the seed fits an ellipse
8	Solidity	The ratio of area over convex area

However, in this phase, several seeds cannot be analyzed automatically (e.g., incorrect seed boundary result in abnormally large area, perimeter, unable to compute skewness and kurtosis). Invalid seed images will be removed as an attempt to recover these images will take more resources than desired. In total, 4400 seed images for spinach and 5250 seed images for water spinach are accepted. The data is now ready for machine learning models with 8 features and the label is either 0 (failed germination) or 1(success germination).

3.3. Training and evaluation of machine learning algorithms

To train and validate a machine learning algorithm, dataset is often divided into two subsets: 75% datapoints go to training set and 25% datapoints go to validation set. Machine learning algorithm is then selected, trained, and fine-tuned on training set. The output of a trained machine learning algorithm is called a model. The model is then validated using the validation set as well as technique such as k-fold cross validation.
In this study, the purpose of the machine learning model is to predict whether a seed will be germinated successfully or not given its image at the beginning of the germination phase. This is a classification problem of two classes: 1 – success germination and 0 – fail germination. Multiple classification machine learning algorithms are investigated to find the optimal one including Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF). These algorithms are known for success in complex classification problems. To determine the nearest optimal configuration for each algorithm, a grid-search [23] is employed on key hyper-parameters.
For ANN, the selected hyper-parameters are shown in Table 3.

Table 3: ANN optimal hyper-parameters

ID	Name	Specification
1	Input layer	8 neurons
2	Output layer	1 neuron (0 or 1)
3	Hidden layer	20 neurons
4	Activation function	Relu
5	Learning rate	0.001
6	Solver	Adam

For SVM, the selected hyper-parameters are shown in Table 4.

Table 4: SVM optimal hyper-parameters

ID	Name	Specification
1	Core function	"rbf"
2	Gamma	0.01
3	Coeficient	0.6
4	C constant	1

And lastly, for RF, the selected hyper-parameters are shown in Table 5.

Table 5: RF optimal hyper-parameters

ID	Name	Specification
1	Number of Decision Tree	100
2	Split function	"Gini"
3	Max-depth	Unlimited
4	Feature separator	"sqrt"

4. Results and Evaluation

Both datasets are divided into 75-25 for training and validation and then applied K-fold cross validation technique for generalization report on models' performance. K is chosen to be 10 as it is sufficient for a generalization report.

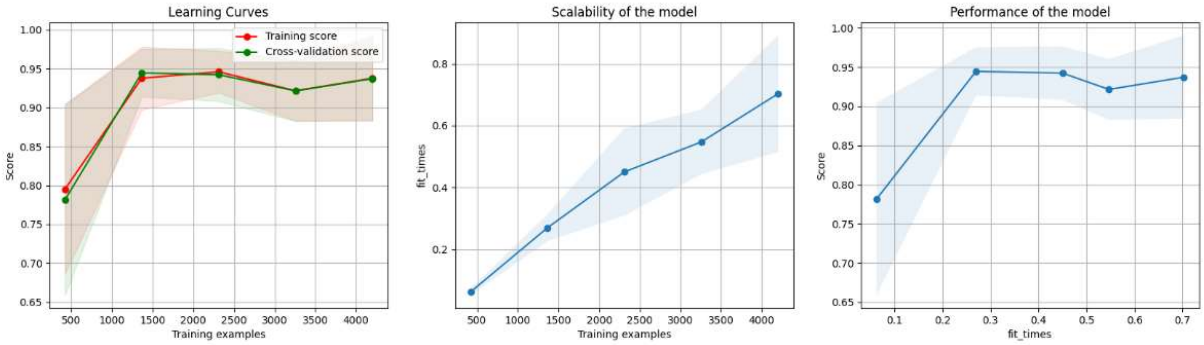

Figure 8: ANN model performance report with 10-fold cross validation on water spinach dataset.

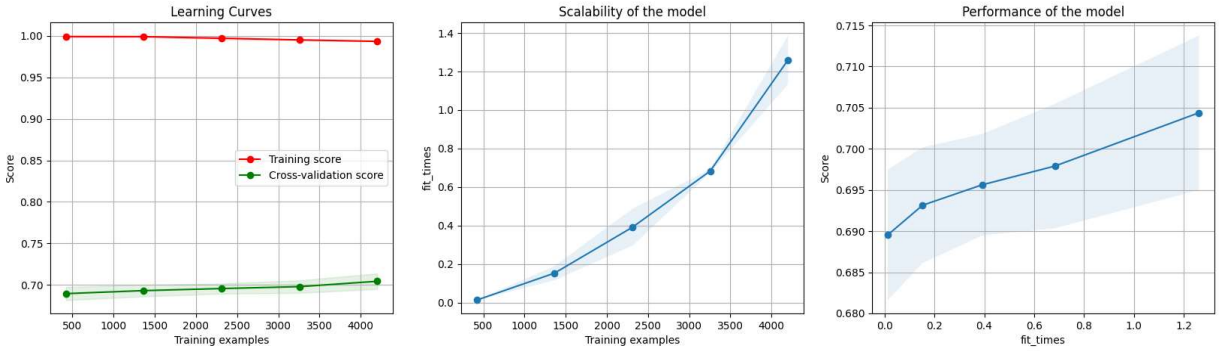

Figure 9: SVM model performance report with 10-fold cross validation on water spinach dataset.

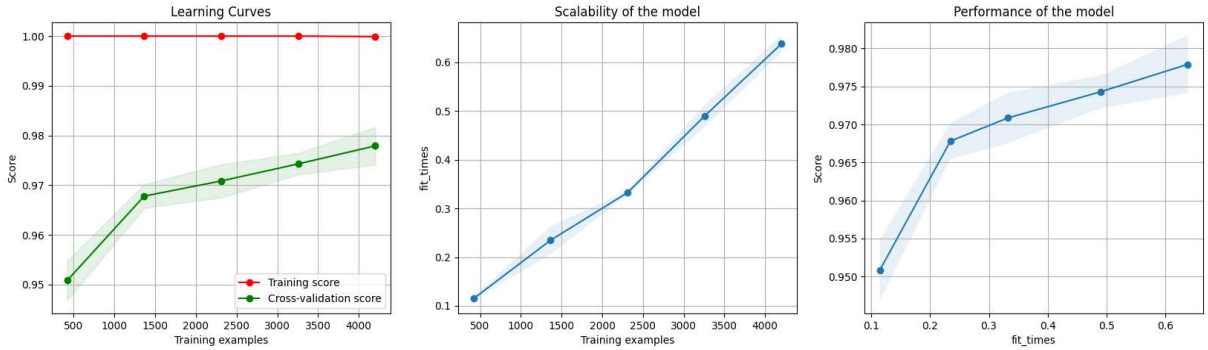

Figure 10: RF model performance report with 10-fold cross validation on water spinach dataset.

For ANN, the model shows a 94.8% classification accuracy on water spinach dataset after running 10-fold cross validation. There is no indication of overfitting as both training and cross-validation scores are closely followed by each other in learning curves. The scalability of the model seems to be linear which means more datapoints will only increase

training time in linear order. The performance of the model graph, however, suggests that more data is not likely to increase the model accuracy.

For SVM, the model performed exceptionally well on training data with 99% accuracy on water spinach dataset. However, there is a clear gap between cross validation score and training score which indicates a likelihood of overfitting. That is the model is likely to perform badly with new unencountered data. The model also likely has an exponential order growth of fit time with more training data and does not seem to improve accuracy significantly if more data is available.

For RF model on water spinach dataset, the model has a close to perfect 100% accuracy on training set. However, the indication for overfitting is clear with a gap between training score and validation score. The model fit time is likely to be linear if more data is introduced and it also shows that more data can enhance the model's performance.

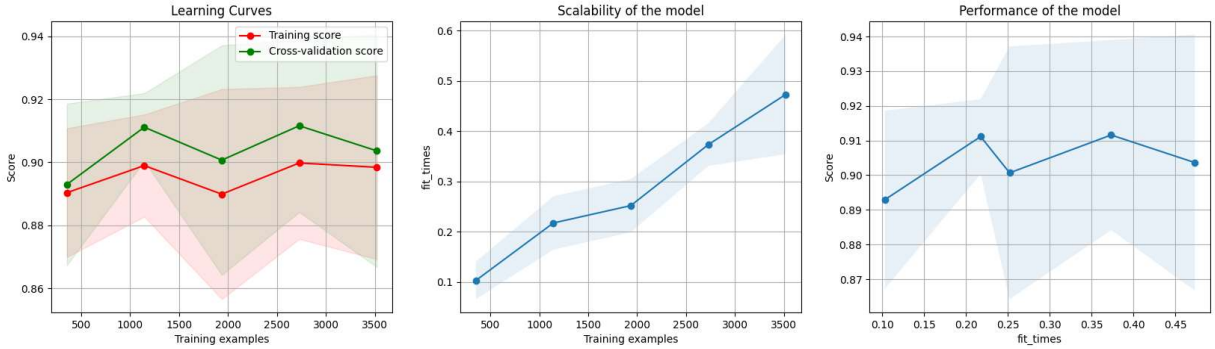

Figure 11: ANN model performance on the spinach dataset.

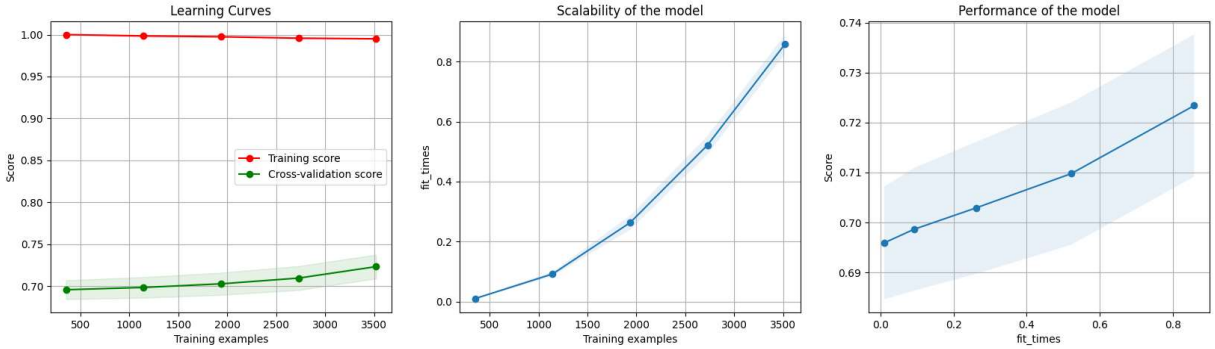

Figure 12: SVC model performance on the spinach dataset.

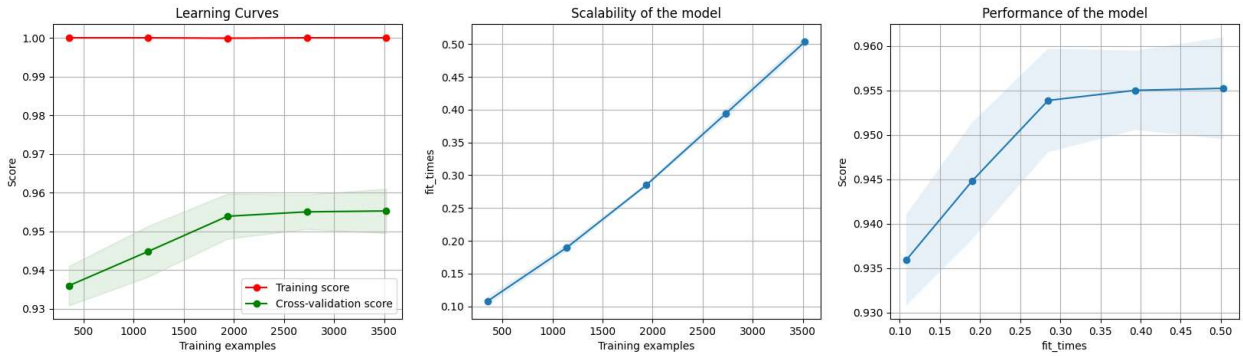

Figure 13: RF model performance on the spinach dataset.

For the spinach dataset, the ANN and SVC model both display similar characteristics to its respective performance on the water dataset (Fig. 11, Fig. 12). The accuracy for ANN is even dropped as low as 90% in the spinach dataset.

When applying the same RF model for the spinach dataset (with 4400 datapoints), the result is promising with 95.6% classification accuracy with 10-fold cross validation (Fig. 13). However, it is shown that the model is not likely to improve after just 2000 training examples.

The comparison between models' performance for water spinach dataset is shown in Table 6.

Table 6: The comparison of models' performance for water spinach dataset

ID	Model	Training score	Validation score
1	ANN	0.94	0.94
2	SVC	0.99	0.71
3	RF	1	0.95

The comparison between models' performance for spinach dataset is shown in Table 7.

Table 7: The comparison of models' performance for spinach dataset

ID	Model	Training score	Validation score
1	ANN	0.89	0.90
2	SVC	0.99	0.73
3	RF	1	0.95

Compared to the state-of-the-art results from deep learning approaches with more than 98% accuracy as in [23], the method is not out-performed the best predictor out there. However, in terms of learning more insight about the relevant factors which contribute to the seed germinability, this study clearly shows relevant factors that can be used in multiple aspects of seeds production industry.

5. Conclusion

In this article, an accurate machine learning approach to predict seed germination is introduced. The study proposes to use 8 features extracted from a non-destructive RGB image of seeds to determine whether the seed is capable of germinating in the future.

Three classification machine learning algorithms are chosen to perform and compare to each other and a validation score of 0.95 is reached with Random Forest model for both datasets with 8 carefully selected and extracted features from seeds' RGB images.

Although the result is promising, it is outperformed by the state-of-the-art deep learning approach by 0.03 point. However, the study provides much more insight into relevant factors which contribute to the germination process.

Acknowledgement

This work was supported by the Vietnam Ministry of Education and Training under project grant number B2020-BKA-12.

References

[1] Kameswara Rao, N., Dulloo, M.E. & Engels, J.M.M. (2017) *A review of factors that influence the production of quality seed for long-term conservation in genebanks*. Genet Resour Crop Evol 64, 1061–1074

[2] Joosen RV, Kodde J, Willems LA, Ligterink W, van der Plas LH, Hilhorst HW (2010). *GERMINATOR: a software package for high-throughput scoring and curve fitting of Arabidopsis seed germination*. Plant J. 2010;62:148–59.

[3] Hoffmaster AF, Xu L, Fujimura K, Bennett MA, Evans AF, McDonald MB (2005) *The Ohio State University seed vigor imaging system (SVIS) for soybean and corn seedlings*. Seed Technol. 2005;27:7–24.

[4] Škrubej U, Rozman Č, Stajnko D. (2015) *Assessment of germination rate of the tomato seeds using image processing and machine learning.* Eur J Hortic Sci. 2015;80:68–75

[5] Nguyen TT, Hoang VN, Le TL, Tran TH, Vu H. (2018) *A vision based method for automatic evaluation of germination rate of rice seeds.* In: 1st international conference on multimedia analysis and pattern recognition (MAPR).

[6] Boelt, B. et al. (2018) *Multispectral imaging–a new tool in seed quality assessment?.* Seed Sci. Res. 28, 222–228

[7] Xia, Y., Xu, Y., Li, J., Zhang, C. & Fan, S. (2019) *Recent advances in emerging techniques for non-destructive detection of seed viability: a review*. Artif. Intell. Agric. 1, 35–47.

[8] ElMasry, G., Mandour, N., Al-Rejaie, S., Belin, E. & Rousseau, D. (2019) Recent applications of multispectral imaging in seed phenotyping and quality monitoring—an overview. Sensors 19, 1090.

[9] LeCun, Y., Bengio, Y. & Hinton, G. (2015) *Deep learning*. Nature 521, 436–444.

[10] Medjahed, Seyyid Ahmed. (2015). *A Comparative Study of Feature Extraction Methods in Images Classification. International Journal of Image, Graphics and Signal Processing*. 7. 16-23. 10.5815/ijigsp.2015.03.03.

[11] Sairi, S.A.M. & Mustafa, S. (2020) *Comparative study of three rice brands' quality through measuring broken rice percentage using Sortex A ColorVision (Buhler) Optical Sorters*. International Conference of Sustainability Agriculture and Biosystem (IOP Publishing, West Sumatera, Indonesia, 2020)

[12] Davis, B. I. et al. (2021) *Measurements of high oleic purity in peanut lots using rapid, single kernel near-infrared refectance spectroscopy*. J. Am. Oil Chem. Soc. 98, 621–632.

[13] Bruggink, H. & Van Duijn, B. (2017) *X-ray based seed analysis.* Seed Test. Int. 153, 45–50.

[14] de Medeiros, A. D. et al. (2020) *Interactive machine learning for soybean seed and seedling quality classifcation.* Sci. Rep. 10, 11267.

[15] ISTA. (2015) *The germination test*. Int Rules Seed Test.

[16] Marcos Filho J, Marcos FJ. Seed vigor testing: an overview of the past, present and future perspective. Sci Agric Scientia Agricola. 2015;72:363–74

[17] Bruggink, H. & Van Duijn, B. (2017) *X-ray based seed analysis*. Seed Test. Int. 153, 45–50.

[18] Ahmed, M. R. et al. (2020) *Classifcation of watermelon seeds using morphological patterns of X-ray imaging: a comparison of conventional machine learning and deep learning*. Sensors 20, 6753.

[19] Galletti, P.A. et al. (2020) *Integrating optical imaging tools for rapid and non-invasive characterization of seed quality: Tomato (Solanumlycopersicum L.) and carrot (Daucus carota L.) as study cases*. Front. Plant Sci. 11, 577851.

[20] Kautzman, M. E., Wickstrom, M. L. & Scott, T. A. (2015) *The use of near infrared transmittance kernel sorting technology to salvage high quality grain from grain downgraded due to Fusarium damage*. Anim. Nutr. 1, 41–46.

[21] Jia, S. et al. (2016) *Feasibility of analyzing frost-damaged and non-viable maize kernels based on near infrared spectroscopy and chemometrics*. J. Cereal Sci. 69, 145–150.

[22] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, (2011) *Non-Local Means Denoising*, Image Processing Online, 1, pp. 208–212.

[23] Jiménez, Á.B., Lázaro, J.L., Dorronsoro, J.R. (2007). *Finding Optimal Model Parameters by Discrete Grid Search*. In: Corchado, E., Corchado, J.M., Abraham, A. (eds) Innovations in Hybrid Intelligent Systems. Advances in Soft Computing, vol 44. Springer, Berlin, Heidelberg.

[24] Genze, N., Bharti, R., Grieb, M. et al. (2020) *Accurate machine learning-based germination detection, prediction and quality assessment of three grain crops*. Plant Methods 16, 157.