

FT NIR 분광법 및 이진분류 머신러닝 방법을 이용한 소나무 종자 발아 예측

김용율¹ · 구자정^{1*} · 구다은¹ · 한심희¹ · 강규석²

¹국립산림과학원 생명정보연구과, ²서울대학교 농림생물자원학부

Prediction of Germination of Korean Red Pine (*Pinus densiflora*) Seed using FT NIR Spectroscopy and Binary Classification Machine Learning Methods

Yong-Yul Kim¹, Ja-Jung Ku^{1*}, Da-Eun Gu¹, Sim-Hee Han¹ and Kyu-Suk Kang²

¹Forest Bioresources Department, National Institute of Forest Science, Suwon 16631, Korea

²Department of Agriculture, Forestry and Bioresources, Seoul National University, Seoul 08826, Korea

요약: 본 연구에서는 -18°C 및 4°C에서 18년간 저장된 소나무 종자 963개에 대해 FT NIR 스펙트럼을 조사하여 7개 머신러닝 방법(XGBoost, Boosted Tree, Bootstrap Forest, Neural Networks, Decision Tree, Support Vector Machine, PLS-DA)을 이용한 종자발아 예측모델을 만들고, 그 성능을 비교하였다. XGBoost 및 Boosted Tree 모델의 예측성능이 가장 우수하였으며, 정확도, 오분류율 및 AUC 값은 각각 0.9722, 0.0278, 0.9735과 0.9653, 0.0347, 0.9647이었다. 2개 모델에서 종자발아 유무를 예측하는 데 있어 상대적 중요도가 높았던 54개 파수 변수들에 대한 파장대는 크게 6개(811~1,088 nm, 1,137~1,273 nm, 1,336~1,453 nm, 1,666~1,671 nm, 1,879~2,045 nm, 2,058~2,409 nm) 그룹으로 나눌 수 있었으며, 방향족 아미노산, 셀룰로스, 리그닌, 전분, 지방산 및 수분과 관련된 것으로 추정되었다. 이상의 결과를 종합할 때, 본 연구에서 얻어진 FT NIR 스펙트럼 데이터와 2개의 머신러닝 모델은 소나무 저장종자의 발아 유무를 정확도 96% 이상으로 예측할 수 있기에 장기저장 종자 유전자원의 비파괴적 활력검정에 유용하게 활용될 수 있을 것으로 생각된다.

Abstract: In this study, Fourier-transform near-infrared (FT-NIR) spectra of Korean red pine seeds stored at -18°C and 4°C for 18 years were analyzed. To develop seed-germination prediction models, the performance of seven machine learning methods, namely XGBoost, Boosted Tree, Bootstrap Forest, Neural Networks, Decision Tree, Support Vector Machine, PLS-DA, were compared. The predictive performance, assessed by accuracy, misclassification, and area under the curve (0.9722, 0.0278, and 0.9735 for XGBoost, and 0.9653, 0.0347, and 0.9647 for Boosted Tree), was better for the XGBoost and decision tree models when compared with other models. The 54 wave-number variables of the two models were of high relative importance in seed-germination prediction and were grouped into six spectral ranges (811~1,088 nm, 1,137~1,273 nm, 1,336~1,453 nm, 1,666~1,671 nm, 1,879~2,045 nm, and 2,058~2,409 nm) for aromatic amino acids, cellulose, lignin, starch, fatty acids, and moisture, respectively. Use of the NIR spectral data and two machine learning models developed in this study gave >96% accuracy for the prediction of pine-seed germination after long-term storage, indicating this approach could be useful for non-destructive viability testing of stored seed genetic resources.

Key words: FT NIR, machine learning, binary classification model, *Pinus densiflora*, seed germination

서론

종자은행을 이용한 식물종의 현지외보존은 비록 적용

대상 종이 한정적이지만, 자연재해로부터 안전하게 자원을 보존할 수 있는 장점이 있다(Food and Agriculture Organization of the United Nations, 2014). 그러나 대다수 산림식물 종의 종자는 저장기간이 길어질수록 종자활력이 감소되기에 주기적인 활력검사가 필요하다(Kim et al., 2012). 일반적으로 종자활력은 발아실험 및 TTC (Triphenyl tetrazolium chloride) 검사에 의해 수행되고 있으나 종자를

* Corresponding author

E-mail: jajungku@korea.kr

ORCID

Ja-Jung Ku  https://orcid.org/0009-0006-3325-4500

소모해야 하며, 결과 확인까지 시간과 인력이 소요되는 단점이 있다. 근적외선에 대한 분광 차이를 이용하여 화합물의 정량적 및 정성적 특성을 분석하는 FT NIR 분광법(Fourier transform near infrared spectroscopy)은 비파괴적이며, 소량의 시료만이 필요하고, 분석 속도가 빠르다는 장점이 있어 농식품의 품질검사(Chen et al., 2021; Shenk et al., 2007) 뿐만 아니라 최근에는 종자품질 및 활력검사(Lestander and Oden, 2002; Qiu et al., 2018; Shetty et al., 2011) 등에도 이용되고 있다.

부분최소제곱 판별분석(PLS-DA, partial least squares discriminant analysis)은 연속형 측정값인 FT NIR 스펙트럼을 설명변수로 사용하여 종자활력 여부와 같은 명목형 반응변수에 대한 이항분류(binary classification) 모델링에 자주 사용되어 왔다(Daneshvar et al., 2015; Tigabu et al., 2020; Tigabu and Odén 2003; Xia et al., 2019). 그러나 변수들 관계가 비선형(non-linear)인 경우, 모델의 예측 성능이 낮아질 수 있고, 모델의 전체변동을 설명하기 위해 다수의 잠재변수(latent variable)가 사용될 경우, 변수 관계에 대한 해석이 어려워질 수 있다(Mo et al., 2020; Ruiz-Perez et al., 2020; Tian et al., 2023). 비선형 관계의 이항분류 모델링을 위한 머신러닝 방법으로는 Decision Tree, 이를 더 발전시킨 Bootstrap Forest, Boosted Tree 및 XGBoost 등이 있으며, 또한 Support Vector Machine과 Neural Networks도 사용되고 있다(Chen and Guestrin, 2016; Kumari and Srivastava, 2017; Narassiguin et al., 2016; Rocha et al., 2020). 머신러닝 방법은 각각의 고유 알고리즘에 따른 장단점이 있고, 데이터의 구조와 변수간 관계 등에 따라 모델의 예측 성능이 달라지기 때문에 어떤 방법이 가장 좋은 것인지를 단정하기 어렵다. 이에 따라 최근에는 다양한 머신러닝 방법을 동시에 사용하여 이 중에서 가장 우수한 예측 성능의 모델을 확립하려는 연구들이 종자품질 및 활력평가 분야에서 자주 시도되고 있다(Liu et al., 2021; Sampaio and Brites, 2021; Wang et al., 2021).

국립산림과학원은 산림생명자원 책임기관으로서 2020년말 현재 74과 432종의 종자 19,902점을 저온 보존하고 있는데, 소나무(*Pinus densiflora Siebold & Zucc.*)의 경우 보존 종자의 19.3%가 20년 이상 저장된 것이어서 활력평가가 필요한 상황이다(Kim et al., 2020). 가급적 보존 종자의 손실없이 단기간에 FT NIR 분광법으로 활력평가를 수행하기 위해서는 높은 예측 성능의 모델이 해당 수종마다 확립되어야 하나, 국내에서는 일본잎갈나무(*Larix kaempferi* (Lamb.) Carrière)와 편백(*Chamaecyparis obtusa* (Siebold & Zucc.) Endl.) 등 2개 수종에 대한 연구만이 보고되어 있을 뿐이다(Mukasa et al., 2018; Mukasa et al., 2019).

이에 본 연구는 국립산림과학원에서 18년간 저장되어 있던 소나무 종자에 대해 FT NIR 스펙트럼 측정과 발아 실험을 수행하고, 3개의 전처리(pre-processing) 방법에 의해 변환 스펙트럼 데이터를 얻은 후, 7개의 머신러닝 방법을 이용하여 종자발아 유무를 정확히 예측할 수 있는 모델을 개발하고자 수행되었다. 보다 세부적으로는 (1) 모델의 예측 성능을 향상시키는 NIR 스펙트럼 데이터의 전처리(pre-processing) 방법을 확인하고, (2) 가장 우수한 예측 성능을 보이는 머신러닝 모델을 확립하며, (3) 모델에서 상대적 중요도가 높은 변수와 관련된 화합물을 추정하고자 하였다.

재료 및 방법

1. 종자 공시재료

국립산림품종관리센터 안면소재 소나무 채종원에서 2003년에 채취되어 국립산림과학원 산림생명자원연구부의 종자저장고에 2021년까지 보존되어 왔던 소나무 2개 클론(강원11호 및 경북40호)의 품대종자 seedlots을 본 연구에 사용하였다. 이들 종자는 탈종, 정선 및 육안식별 작업을 거쳐 선별된 충실종자(sound seed)이며, 함수율 5~7%로 건조되어 밀봉된 뒤, -18°C(습도 30%) 및 4°C(40%)에서 지난 18년간 저장되어 있었다.

2. FT NIR 스펙트럼 측정

클론마다 각 저장온도별로 300립의 종자를 무작위로 샘플링하여 총 1,200립의 종자를 FT NIR 스펙트럼 측정에 사용하였다. Quartz beam splitter와 Integrating sphere가 장착된 MPA II (Bruker Karlsruhe, Germany)를 이용하여 종자의 FT NIR 확산반사(diffuse reflectance) 스펙트럼을 측정하였다. 측정 범위는 12,481~3,995 cm^{-1} , 측정 간격은 15 cm^{-1} 로 하였으며, 64회 반복 측정(속도: 10 kHz)하여 평균값을 1회 측정에 대한 데이터로 하였고, 이는 background 및 종자 측정 모두에 동일하게 적용되었다. 각 종자마다 3회 반복 측정하여 평균값을 최종 스펙트럼 데이터로 하였다. 또한 시료측정 용기를 사용하여 모든 종자들에 대한 측정이 가급적 동일한 위치에서 이루어지도록 하였다. 각 종자의 FT NIR 확산반사 스펙트럼 데이터에 대한 기준선 보정(baseline correction)과 흡광(absorbance) 데이터로의 변환은 OPUS Base Package (Version 8.0) 소프트웨어를 사용하여 수행하였다. FT NIR 스펙트럼 측정을 한 모든 종자는 개별 고유번호를 부여하여 이후의 분석에 사용하였다.

3. 종자발아 실험

FT NIR 스펙트럼 측정이 완료된 1,200립의 종자를 1%

agar 배지가 들어있는 petri dish에 25립씩 치상하였다. Petri dish안의 종자 위치를 기록하여 각 종자의 발아여부 결과를 해당 종자의 FT NIR 스펙트럼 데이터에 정확히 연계되도록 하였다. Petri dish를 25°C의 성장상에 옮기고, 80일 동안 16시간의 광 처리와 8시간의 암 처리를 하여 종자 발아를 유도하였다. 80일 동안 유근(radicle)이 2 mm 이상 자란 것을 “발아” 종자로 그 반대의 경우는 “미발아” 종자로 규정하였다. 발아 과정에서 곰팡이가 번식된 것을 제외한 총 963개 종자에 대한 발아 여부 데이터를 얻어 머신러닝 분석에서 반응변수 y 의 값으로 사용하였다.

4. FT NIR 측정 데이터의 변환

963개 종자에 대한 FT NIR 스펙트럼 데이터에 대해 JMP Pro® 17 프로그램의 “Functional Data Explorer” 기능(JMP Statistical Discovery LLC, 2022)을 이용하여 Savitzky-Golay smoothing 알고리즘(Savitzky and Golay, 1964)에 의한 2차 다항식(이하 “SG 2차항”으로 표기) 변환, SG 2차항 변환 데이터에 대한 1차 미분(이하 “SG 2차항 + 1차 미분”으로 표기)과 2차 미분(이하 “SG 2차항 + 2차 미분”으로 표기) 변환을 하였다. 최종적으로 모두 4개의 FT NIR 스펙트럼 데이터(미변환 원래 데이터 1, 변환 데이터 3)를 이후의 머신러닝 분석에 사용하였다. 각 데이터에서 열변수의 이름(column name)은 “ab0000” 형식으로 부여하였는데, 예로써 “ab7004”는 7,004 cm^{-1} 의 파수 변수를 의미한다.

5. 머신러닝 분석

머신러닝 분석에 있어 데이터 분할, 모델 선별, 예측 성능 평가는 기본적으로 JMP® Pro 17 프로그램을 이용하여 수행되었다. 비복원 무작위 층화추출(stratified random sampling without replacement) 방법에 따라 4개 FT NIR 스펙트럼 데이터를 훈련(70%) 및 테스트(30%) 세트로 분할하였다. 이때, 발아유무, 저장온도 및 seedlots의 비율이 데이터 분할에 반영되도록 하였다.

JMP® Pro 17 프로그램의 “Model Screening” 기능(JMP

Statistical Discovery LLC, 2022)을 사용하여 4개 FT NIR 스펙트럼 데이터의 훈련세트를 7개 머신러닝 방법(XGBoost, Boosted Tree, Bootstrap Forest, Neural Networks, Support Vector Machine, Decision Tree, PLS-DA)에 적용하고, 7겹 교차검증(7-fold cross validation)을 수행하였다. 훈련 세트의 검증데이터 세트(validation data set)에 대한 AUC (area under ROC curve), 오분류율(misclassification) 및 R^2 값을 산출하여 데이터 종류 및 머신러닝 방법에 따른 학습모델의 예측성능을 비교하였다.

가장 우수한 예측성능을 보이는 FT NIR 스펙트럼 데이터와 상위 4개의 머신러닝 모델(full model)을 선택하고, 테스트 세트에 적용하여 각 모델의 정확도(accuracy), 오분류율(misclassification), 민감도(sensitivity, recall), 특이도(specificity) 및 정밀도(precision)를 산출하여 모델의 예측 성능을 비교하였다. 정확도, 오분류율, 민감도, 특이도 및 정밀도는 종자 발아의 경우를 positive, 종자 미발아 경우를 negative로 할 때, 발아 및 미발아된 실제 종자의 비율과 모델에 의해 예측된 발아 및 미발아 종자의 비율에 대한 2×2 혼동행렬(confusion matrix)를 만들어(Table 1) 다음의 식에 따라 계산되었다.

$$\text{정확도 (accuracy)} = \frac{TP+TN}{TP+FN+TN+FP}$$

$$\text{오분류율 (misclassification)} = 1 - \text{정확도}$$

$$\text{민감도 (sensitivity, recall)} = \frac{TP}{TP+FN}$$

$$\text{특이도 (specificity)} = \frac{TN}{TN+FP}$$

$$\text{정밀도 (precision)} = \frac{TP}{TP+FP}$$

예측성능 상위 4개 모델 각각에 대하여 상대적 중요도(relative variable importance)가 높은 변수만을 골라 축소 모델(reduced model)을 만들었다(JMP Statistical Discovery LLC, 2022). 즉, 해당 변수 자체와 다른 변수를 고려한 분산을 서로 합하여 전체 모델의 분산으로 나눈 총효과(total

Table 1. Confusion matrix for evaluation of model classification performance.

Classification		Predicted value	
		Positive (germination)	Negative (non-germination)
Actual value	Positive (germination)	TP (True Positive)	FP (False Positive)
	Negative (non-germination)	FN (False Negative)	TN (True Negative)

TP (True Positive): Number of germinated seeds correctively classified as the germinated seeds by the model.

FP (False Positive): Number of germinated seeds mis-classified as the non-germinated seeds by the model.

FN (False Negative): Number of non-germinated seeds mis-classified as the germinated seeds by the model.

TN (True Negative): Number of non-germinated seeds correctively classified as the non-germinated seeds by the model.

effect) 값이 0.01 이상인 변수만을 선택하여 훈련세트를 만들고, 7겹 교차검증을 통해 모델을 학습시킨 뒤, 테스트 세트에 적용하여 예측성능이 가장 우수한 상위 2개 모델을 최종적으로 선택하였다.

6. 최적 모델의 파장변수 비교 및 관련 화합물 추정

예측성능이 가장 우수한 2개 모델의 파수 변수에 대한 파장(wavelength, nm) 영역대를 *Pinus patula*, *P. massoniana*, *P. elliotii*, *P. taeda* 종자에 대한 NIR 스펙트럼 연구결과(Tigabu, 2003; Tigabu and Odén, 2003; Tigabu et al., 2019)와 목재성분 NIR 스펙트럼 연구결과(Schwanninger et al., 2011) 그리고 Workman and Weyer(2012)의 자료와 비교하여 그룹화하였고, 두 모델 공통의 파수 변수를 확인하였으며, 그룹별 파수 변수와 관련된 화합물을 추정하였다.

결과 및 고찰

1. 종자발아 및 FT NIR 스펙트럼

963개 종자 중에서 발아종자는 540개(56.1%), 미발아는

423개(43.9%)였다(Table 2). Fisher의 정확 검정(Fisher's Exact Test) 결과, 저장온도에 따른 발아종자의 비율은 동일하지 않았으며(양측검정 $p = < 0.0001$), 4°C에서의 발아종자 비율이 -18°C 보다 더 낮았다. 이와 달리 Seedlots에 따른 발아종자의 비율은 유사하였다(양측검정 $p = 0.1527$).

963개 종자의 FT NIR 스펙트럼 프로파일은 1,440 nm와 1,923 nm 파장 영역대에서 정점을 보였는데[Figure 1(A)], 이는 *Pinus massoniana*, *P. elliotii* 및 *P. taeda*에서의 정점 파장대인 1,450 nm 및 1,936 nm와 유사하였다(Tigabu et al., 2019). 발아종자 그룹의 평균 흡광도는 모든 파장에서 미발아 그룹에 비해 더 높았으나[Figure 1(B)], 실제로는 전체 파장 영역대에서 두 그룹의 종자 흡광도가 그룹 간에 뚜렷한 차이를 보이지 않고 서로 중첩되어 있어 단순히 평균 흡광도 스펙트럼만을 이용하여 종자발아 유무를 판별하기는 사실상 불가능하였다.

2. FT NIR 스펙트럼 데이터 종류에 따른 7개 머신러닝 방법의 예측성능 비교

PLS-DA 등 7개 머신러닝 방법을 4개 FT NIR 스펙트럼

Table 2. Number of germinated and non-germinated seeds by seedlots and storage temperature.

Seedlots (clone name)	Storage temperature	Number of germinated seeds	Number of non-germinated seeds	Total
GB40 (Gyeongbuk 40)	-18°C	210	54	264
	4°C	57	135	192
	Subtotal	267	189	456 (47.4%)
GW11 (Gangwon 11)	-18°C	189	78	267
	4°C	84	156	240
	Subtotal	273	234	507 (52.6%)
Total		540 (56.1%)	423 (43.9%)	963

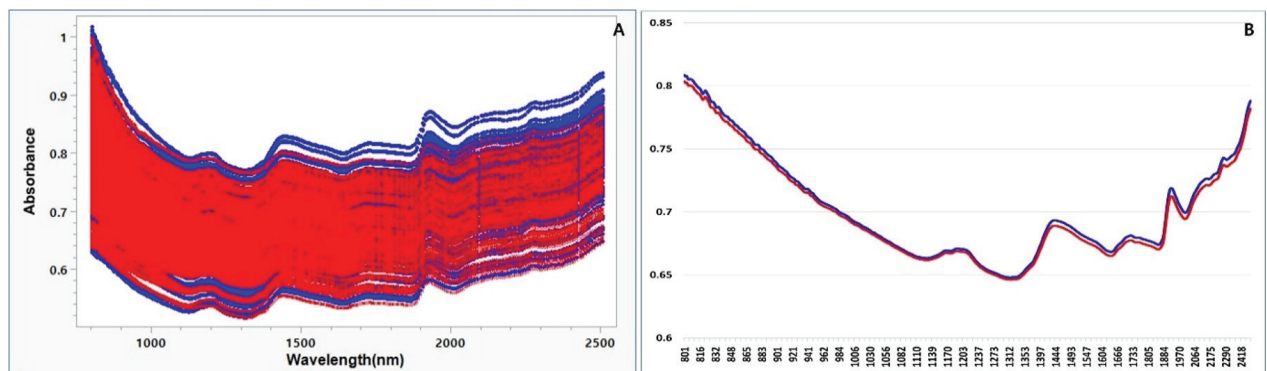


Figure 1. FT NIR spectral profile(A) and group mean absorbance spectral profile(B) for germinated (blue line) and non-germinated (red line) seeds.

데이터에 적용하여 7-fold 교차검증을 수행한 결과, 4개 머신러닝 방법(XGBoost, Boosted Tree, Bootstrap Forest, Neural Networks)의 예측성능(AUC: 0.879 ~ 0.942, 오분류율: 0.122 ~ 0.196, R^2 : 0.339 ~ 0.542)이 4개 데이터 모두에서 나머지 3개 방법(Decision Tree, Support Vector Machine,

PLS-DA) 보다 더 우수하였다(Table 3).

데이터의 종류에 따른 7개 머신러닝 방법의 평균 예측 성능을 비교한 결과, SG 2차항 + 1차미분 변환 데이터가 AUC 값 및 오분류율 값에서 각각 평균 0.876 및 0.180으로 가장 우수하였고, R^2 값에서는 오히려 SG 2차항 변환

Table 3. Predictive performance of seven machine learning methods on four FT NIR spectral data resulting from 7-fold cross-validation.

FT NIR spectral data	Machine learning models	AUC	Misclassification	R^2
Raw data	XGBoost	0.964	0.089	0.661
	Bootstrap Forest	0.940	0.123	0.474
	Boosted Tree	0.937	0.133	0.533
	Neural Networks	0.904	0.191	0.422
	Decision Tree	0.821	0.227	0.250
	Support Vector Machine	0.735	0.338	0.108
	PLS-DA	0.766	0.131	0.406
	Mean (7 models)	0.867	0.176	0.408
	Mean (top 4 models)	0.936	0.134	0.523
	SG (2 nd polynomial) transformed data	XGBoost	0.965	0.081
Bootstrap Forest		0.950	0.115	0.475
Boosted Tree		0.940	0.123	0.581
Neural Networks		0.914	0.167	0.435
Decision Tree		0.850	0.193	0.328
Support Vector Machine		0.734	0.338	0.106
PLS-DA		0.749	0.249	0.371
Mean (7 models)		0.872	0.181	0.424
Mean (top 4 models)		0.942	0.122	0.542
SG (2 nd polynomial) + 1 st derivative transformed data		XGBoost	0.956	0.086
	Bootstrap Forest	0.946	0.123	0.488
	Boosted Tree	0.936	0.124	0.534
	Neural Networks	0.891	0.190	0.395
	Decision Tree	0.843	0.247	0.271
	Support Vector Machine	0.795	0.253	0.227
	PLS-DA	0.762	0.236	0.369
	Mean (7 models)	0.876	0.180	0.413
	Mean (top 4 models)	0.932	0.131	0.507
	SG (2 nd polynomial) + 2 nd derivative transformed data	XGBoost	0.900	0.173
Bootstrap Forest		0.890	0.185	0.353
Boosted Tree		0.880	0.184	0.376
Neural Networks		0.844	0.241	0.277
Decision Tree		0.821	0.259	0.122
Support Vector Machine		0.733	0.287	0.174
PLS-DA		0.770	0.227	0.352
Mean (7 models)		0.834	0.222	0.286
Mean (top 4 models)		0.879	0.196	0.339

데이터가 평균 0.424로 가장 우수하였다(Table 3). 그러나 예측성능 상위 4개의 머신러닝(XGBoost, Boosted Tree, Bootstrap Forest, Neural Networks)만을 고려하여 4개 데이터를 비교하면, SG 2차항 변환 데이터가 가장 우수한 평균 예측성능(AUC: 0.942, 오분류율: 0.122, R^2 : 0.542)을 보였다.

3. 4개 머신러닝 모델의 예측성능 비교

예측성능 상위 4개 머신러닝 모델(full model)을 SG 2차항 변환 데이터의 테스트 세트(288개)에 적합시킨 결과, XGBoost 모델이 가장 우수한 예측성능(정확도: 0.9688, 오분류율: 0.0313, AUC: 0.9661)을 보였다(Table 4). 나머지 모델의 예측성능은 Boosted Tree (0.9201, 0.0799, 0.9176), Bootstrap Forest (0.8889, 0.1111, 0.8801), Neural Networks (0.8125, 0.1875, 0.8086)의 순이었다. 4개 머신러닝 모델의 평균 예측성능은 정확도 0.8976, 오분류율 0.1025, AUC 0.8931로 10개 잠재변수로 확립한 PLS-DA 모델(0.7326, 0.2674, 0.7297) 보다 높았다.

4개 머신러닝 모델의 변수 중에서 총효과가 0.01 이상인 것을 선택(XGBoost: 28개 변수, Boosted Tree: 33개,

Bootstrap Forest: 15개, Neural Networks: 195개)하여 축소 모델(reduced model)을 만들고, 테스트 세트에 적용한 결과, XGBoost 축소모델이 가장 우수한 예측성능(정확도: 0.9722, 오분류율: 0.0278, AUC: 0.9735)을 보였다(Table 4). 4개 축소모델의 평균 예측성능은 정확도 0.9063, 오분류율 0.0938, AUC 0.9039로 10개 잠재변수, 311개 변수로 확립한 PLS-DA 축소모델(0.7431, 0.2569, 0.7381) 보다 높았다.

이상의 결과를 종합하면, XGBoost 및 Boosted Tree 머신러닝 방법을 SG 2차항 변환 FT NIR 스펙트럼 데이터에 적용할 경우, 각각 28개 및 33개의 변수로 정확도 0.9722 및 0.9653(오분류율: 0.0278, 0.0347)의 높은 예측성능을 보이는 예측 모델을 확립할 수 있었다.

4. XGBoost 및 Boosted Tree 축소모델

28개 변수로 확립된 XGBoost 축소모델은 주요 hyperparameter 값이 max depth=6, alpha=0, lamda=1, gamma=0, learning rate=0.1, iterations=30, classification decision threshold=0.5, booster=gmtree (tree-based boosting algorithm) 일 때, 테스트 세트에 대해 정확도 0.9722, 오분류율 0.0278,

Table 4. Comparison of the predictive performance of four machine learning models on testing data.

Model	XGBoost	Boosted Tree	Bootstrap Forest	Neural Networks	4 models mean	PLS-DA	
Full model	No. of variables	530	530	530	-	530	
	Accuracy	0.9688	0.9201	0.8889	0.8125	0.8976	0.7326
	Misclassification	0.0313	0.0799	0.1111	0.1875	0.1025	0.2674
	AUC	0.9661	0.9176	0.8801	0.8086	0.8931	0.7297
	Sensitivity	0.9877	0.9383	0.9506	0.8395	0.9290	0.7531
	Specificity	0.9444	0.8968	0.8095	0.7778	0.8571	0.7063
	Precision	0.9581	0.9212	0.8652	0.8293	0.8934	0.7673
Reduced model	No. of variables	28	33	15	195	-	311
	Accuracy	0.9722	0.9653	0.8576	0.8299	0.9063	0.7431
	Misclassification	0.0278	0.0347	0.1424	0.1701	0.0938	0.2569
	AUC	0.9735	0.9647	0.8514	0.8258	0.9039	0.7381
	Sensitivity	0.9630	0.9691	0.9012	0.8580	0.9228	0.7778
	Specificity	0.9841	0.9603	0.8016	0.7937	0.8849	0.6984
	Precision	0.9873	0.9691	0.8538	0.8424	0.9132	0.7683

Table 5. Comparison of the predictive performance of XGBoost reduced models on training, validation, and testing data.

Data	Sample (N)	Accuracy	Miss-classification	AUC	Sensitivity	Specificity	Precision
Training	579	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000
Validation	96	0.9792	0.0208	0.9924	0.9821	0.9750	0.9821
Testing	288	0.9722	0.0278	0.9735	0.9630	0.9841	0.9873

AUC 0.9735의 매우 높은 예측성능을 보였다(Table 5). 검증 및 테스트 세트 간에 예측성능의 차이가 거의 없어 모델의 과적합 가능성은 없는 것으로 보이며, 이에 따라 본 연구에서 확립된 XGBoost 축소모델은 새로운 데이터에 대해서도 유사한 예측성능을 보일 것으로 생각된다.

XGBoost 축소모델의 28개 변수 중 4개(ab4150, ab6001, ab6881, ab7004)가 0.1 ~ 0.187의 총효과 값을 보여 다른 변수 보다 상대적 중요도가 높았다[Figure 2(A)]. 이들 4개 변수들에 있어 흡광도에 따른 종자발아 유무 예측 확률은 ab7004의 경우 흡광도 0.671 이상, ab6001은 0.6766 이하, ab4150은 0.8167 이상, ab6881은 0.6877 이상으로 될 때, 발아종자로 예측될 확률이 높았다[Figure 2(B)].

33개 변수로 확립된 Boosted Tree 축소모델은 모델의 주요 하이퍼파라미터 값이 number of splits per tree=7, learning rate=0.072, overfit penalty=0.0001, number of tree layers=169, classification decision threshold=0.5일 때, 테스트 세트에 대해 정확도 0.9653, 오분류율 0.0347, AUC 0.9647의 예측성능을 보였다(Table 6). XGBoost 축소모델

과는 달리 테스트 세트이 검증 세트에 비해 더 나은 예측 성능을 보여, 모델이 과소적합된 것으로 보이나 그 차이가 3~5% 범위여서 큰 문제가 없는 것으로 생각된다. 다만, 새로운 데이터에 대해서도 상기 Boosted Tree 축소모델이 유사한 예측성능을 보일 것인가에 대해서는 현재로서는 판단하기 어려우며, 추가적인 분석이 필요하다고 생각된다.

Boosted Tree 축소모델의 33개 변수 중 4개(ab12311, ab10229, ab4196, ab4474)가 0.089 ~ 0.213의 총효과 값을 보여 모델에 대한 상대적 중요도가 나머지 29개 변수보다 높았다[Figure 3(A)]. 4개 변수의 흡광도에 따른 종자발아 유무 예측 확률은 ab12311의 경우 흡광도 0.6769 이상, ab10229는 0.6039 이하, ab4196은 0.7364 이상, ab4474는 0.7376 이하로 될 때, 발아종자로 예측될 확률이 높았다 [Figure 3(B)].

5. FT NIR 파수 변수 비교 및 관련 화합물 추정

XGBoost 축소모델의 28개 변수와 Boosted Tree 축소모델의 33개 변수에 대한 파장 영역대를 *P. patula*, *P. massoniana*,

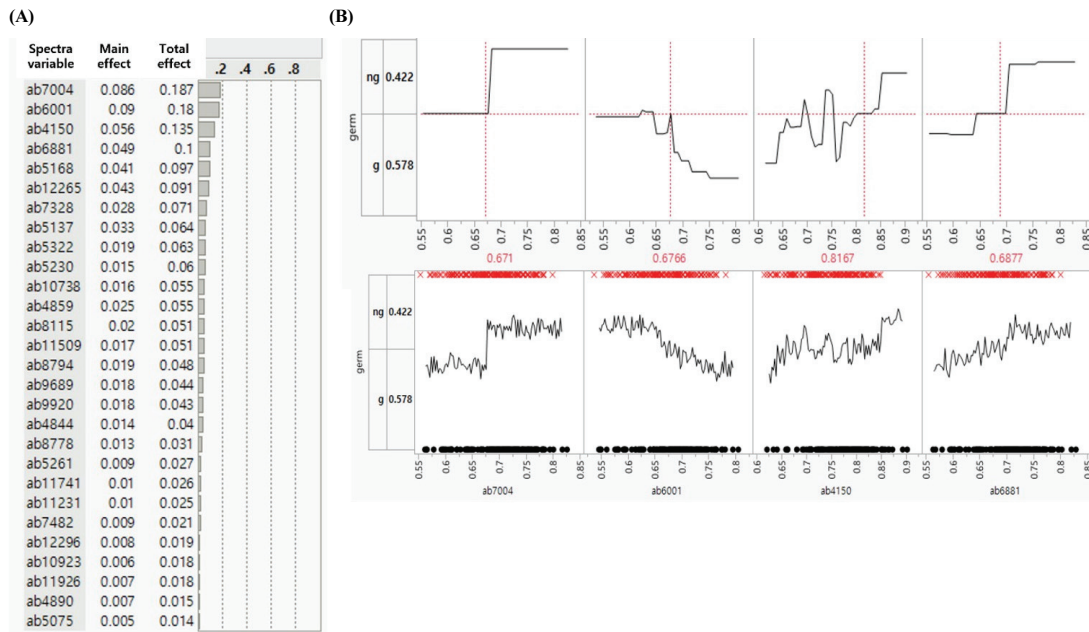


Figure 2. Relative importance of 28 wavenumber variables of the XGBoost reduced model (A) and seed germination prediction profile of each variables (B). The red Xs and black closed circles in the bottom of the Figure 2-B indicate non-germinated and germinated seeds, respectively.

Table 6. Comparison of the predictive performance of Boosted Tree reduced models on training, validation, and testing data.

Data	Sample (N)	Accuracy	Miss-classification	AUC	Sensitivity	Specificity	Precision
Training	579	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000
Validation	96	0.9271	0.0729	0.9931	0.9362	0.9184	0.9167
Testing	288	0.9653	0.0347	0.9691	0.9691	0.9603	0.9691

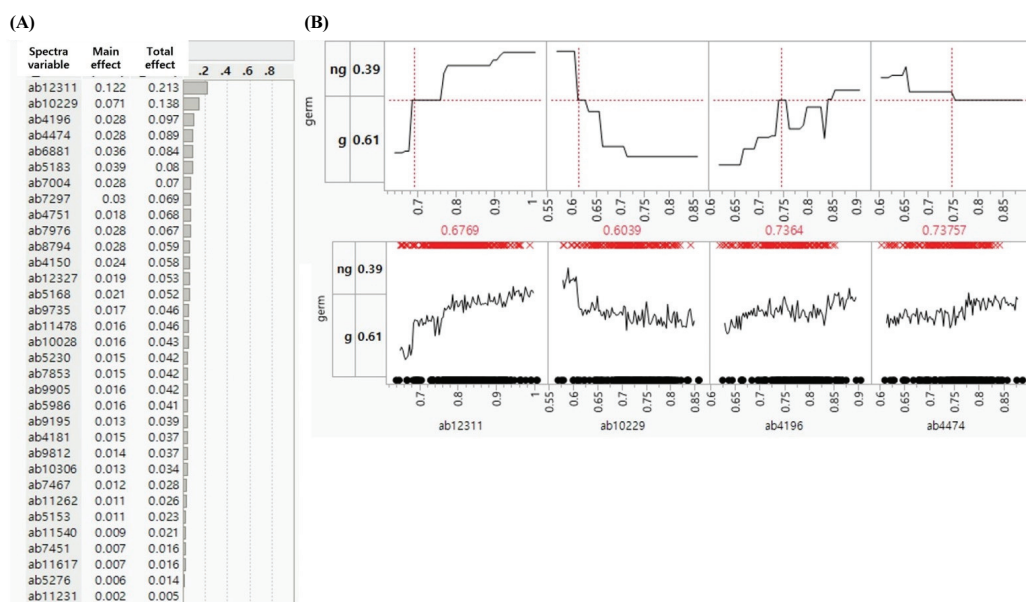


Figure 3. Relative importance of 33 wavenumber variables of the Boosted Tree reduced model (A) and seed germination prediction profile of each variables (B). The red Xs and black closed circles in the bottom of the Figure 2-B indicate non-germinated and germinated seeds, respectively.

Table 7. Fifty-four wavenumber variables aligned by wavelength of the XGBoost and Boosted Tree reduced models.

Group	Wave number variable			XGBoost		Boosted Tree	
	Variable name	Wave number (cm ⁻¹)	Wavelength (nm)	Main effect	Total effect	Main effect	Total effect
Group I (811~1,088 nm)	ab12327	12,327	811			0.0069	0.0278
	ab12311	12,311	812			0.0519	0.1307
	ab12296	12,296	813	0.0092	0.0188		
	ab12265	12,265	815	0.0431	0.086		
	ab11926	11,926	839	0.0074	0.0159		
	ab11741	11,741	852	0.0127	0.0274		
	ab11617	11,617	861			0.0181	0.0712
	ab11540	11,540	867			0.0081	0.0276
	ab11509	11,509	869	0.0183	0.0477		
	ab11478	11,478	871			0.0127	0.0436
	ab11262	11,262	888			0.0034	0.01
	ab11231	11,231	890	0.0118	0.0268	0.0089	0.0326
	ab10923	10,923	915	0.0098	0.0212		
	ab10738	10,738	931	0.0242	0.0573		
	ab10306	10,306	970			0.0085	0.0268
	ab10229	10,229	978			0.0317	0.0966
	ab10028	10,028	997			0.0694	0.1987
	ab9920	9,920	1,008	0.0171	0.0402		
	ab9905	9,905	1,010			0.0128	0.0489
	ab9812	9,812	1,019			0.0099	0.0359
ab9735	9,735	1,027			0.0049	0.0142	
ab9689	9,689	1,032	0.0219	0.0442			
ab9195	9,195	1,088			0.0274	0.0977	

Table 7. (Continued)

Group	Wave number variable			XGBoost		Boosted Tree	
	Variable name	Wave number (cm ⁻¹)	Wavelength (nm)	Main effect	Total effect	Main effect	Total effect
Group II (1,137~1,273 nm)	ab8794	8,794	1,137	0.023	0.0519	0.0146	0.0477
	ab8778	8,778	1,139	0.0159	0.0327		
	ab8115	8,115	1,232	0.0191	0.045		
	ab7976	7,976	1,254			0.0023	0.0084
	ab7853	7,853	1,273			0.0084	0.0261
Group III (1,336~1,453 nm)	ab7482	7,482	1,336	0.0095	0.0212		
	ab7467	7,467	1,339			0.0052	0.0207
	ab7451	7,451	1,342			0.0051	0.0186
	ab7328	7,328	1,365	0.0332	0.0741		
	ab7297	7,297	1,370			0.0154	0.0666
	ab7004	7,004	1,428	0.0895	0.1833	0.0136	0.0423
	ab6881	6,881	1,453	0.0472	0.1000	0.0351	0.1033
Group IV (1,666~1,671 nm)	ab6001	6,001	1,666	0.0884	0.1861		
	ab5986	5,986	1,671			0.0165	0.0618
Group V (1,879~2,045 nm)	ab5322	5,322	1,879	0.0232	0.0606		
	ab5276	5,276	1,895			0.0047	0.0185
	ab5261	5,261	1,901	0.0096	0.0278		
	ab5230	5,230	1,912	0.0176	0.0604	0.0171	0.0494
	ab5183	5,183	1,929			0.0168	0.0473
	ab5168	5,168	1,935	0.0488	0.0986	0.0193	0.0652
	ab5153	5,153	1,941			0.0206	0.0671
	ab5137	5,137	1,946	0.0382	0.0690		
	ab5075	5,075	1,970	0.0056	0.0136		
Group VI (2,058~2,409 nm)	ab4890	4,890	2,045	0.0069	0.0141		
	ab4859	4,859	2,058	0.0291	0.0581		
	ab4844	4,844	2,064	0.0177	0.0376		
	ab4751	4,751	2,104			0.0211	0.0935
	ab4474	4,474	2,235			0.006	0.0226
	ab4196	4,196	2,383			0.023	0.0771
	ab4181	4,181	2,392			0.0075	0.0245
	ab4150	4,150	2,409	0.0567	0.1326	0.0518	0.1597

The variables highlighted in bold are present in both models.

P. elliottii, *P. taeda*에서의 연구결과(Tigabu, 2003; Tigabu and Odén, 2003; Tigabu et al., 2019) 및 Schwanninger et al.(2011)과 Workman and Weyer(2012)의 자료와 비교한 결과, 전체 54개 파장 영역대를 6개 그룹으로 나눌 수 있었으며, 7개 파수 변수(ab11231, ab8794, ab7004, ab6881, ab5230, ab5168, ab4150)가 두 모델에서 공통이었다(Table 7).

Group I은 23개의 변수(XGBoost: 10개, Boosted Tree:

14개)를 포함하고 있으며, 두 모델의 공통변수는 ab11231 (890 nm)로 총효과는 각각 0.0268 및 0.0326이었다. *P. patula* 종자에서는 850~880 nm 및 890~940 nm가 각각 벤젠 및 오일의 C-H 신축진동(stretching vibration)의 3차 배음(third overtone) 영역대, 그리고 1010~1030 nm는 방향족 아미노산(aromatic amino acids, ArNH₂)의 N-H 신축진동의 2차 배음과 C-H 신축진동 및 변형(deformation) 영역

대라고 보고된 바 있다(Tigabu, 2003; Tigabu and Odén, 2003). Group I에서 9개 변수(XGBoost: 5, Boosted Tree: 5, 공통: 1)의 영역대 852~931 nm와 5개 변수(XGBoost: 2, Boosted Tree: 3)의 영역대 1,008~1,032 nm가 이들과 중첩되어 있어 벤젠과 오일, 그리고 방향족 아미노산과 관련된 것으로 생각된다.

Group II에는 5개의 변수(XGBoost: 3개, Boosted Tree: 3개)가 포함되어 있으며, 두 모델의 공통변수는 ab8794 (1,137 nm)로 총효과는 각각 0.0519 및 0.0477이었다. *P. patula* 종자에서 1,100~1,300 nm(중앙값: 1,206 nm)는 CH₃, CH₂ 및 C2H₂ 작용기에 있어 C-H 신축진동 2차 배음 영역대라고 하며(Tigabu and Odén, 2003), 또한 올레산(oleic acid)과 같은 지방산(fatty acids)은 1,180 nm 파장 영역대에서 뚜렷한 피크의 스펙트럼을 보인다고 한다(Sato et al., 1991). Group II의 2개 변수 ab8778(1,139 nm) 및 ab8115(1,232 nm)가 이들과 중첩되어 있어 지방산과 관련된 것으로 생각된다.

Group III에는 7개 파수 변수(XGBoost: 4개, Boosted Tree: 5개)가 포함되어 있으며, 두 모델의 공통변수는 ab7004(1,428 nm) 및 ab6881(1,453 nm)로 총효과는 각각 0.1833 및 0.0423, 0.1000 및 0.1033이었다. 목재의 경우 1,428 nm에서의 스펙트럼이 셀룰로스(cellulose) 및 수분에 있어 O-H 신축진동 1차 배음에 의한 것이라고 보고된 바 있으며(Schwanninger et al., 2011), *Pinus patula*, *P. massoniana*, *P. elliottii*, *P. taeda* 종자에서는 1,450 nm에서의 스펙트럼이 ROH, 단백질, 전분 및 수분 등에 있어 O-H 및 N-H 신축진동 1차 배음과 C-H 결합에 의한 것으로 보고된 바 있다(Tigabu et al., 2019). 이를 고려한다면, Group III의 2개 공통 변수인 ab7004(1,428 nm) 및 ab6881(1,453 nm)은 각각 셀룰로스 및 수분, 그리고 ROH, 단백질, 전분 및 수분과 관련된 것으로 보인다.

Group IV에는 2개의 변수 ab6001(1,666 nm)과 ab5986 (1,671 nm)가 포함되어 있으며, 각 변수의 총효과는 0.1861 및 0.0618이었다. ab6001(1,666 nm) 변수는 spruce 목재(Schwanninger et al., 2011)에 있어 hemicellulose의 CH₃ 작용기의 C-H 신축진동 영역대인 1,666 nm과 일치하였으며, ab5986(1,671 nm)는 리그닌의 C-H 신축진동 1차 배음 영역대인 1,673 nm와 매우 근접한 것이어서 이들 2개 변수는 각각 hemicellulose 및 리그닌과 관련된 것으로 생각된다.

Group V에는 10개의 변수(XGBoost: 7개, Boosted Tree: 5개)가 포함되어 있으며, 두 모델의 공통변수는 ab5230 (1,912 nm) 및 ab5168(1,935 nm)로 총효과는 각각 0.0604 및 0.0494, 그리고 0.0986 및 0.0652이었다. *P. patula* 종자

에서는 1,850 ~ 2,050 nm 영역대에서 1,926 nm를 중심으로 하는 스펙트럼 곡선을 보이는데, 단백질에서의 C=O 신축진동 2차 배음, 전분에서의 O-H 신축진동 및 HOH 변형(deformation) 결합, 그리고 수분에서의 O-H 굽힘진동(bending vibration) 2차 배음 영역대인 것으로 보고된 바 있다(Tigabu and Odén, 2003). 한편, Tigabu et al.(2019)은 *P. massoniana*, *P. elliottii*, 및 *P. taeda*의 종자에서 수분과 관련성이 높은 1,930 nm 파장이 정상적인 발아종자를 배(embryo) 등이 형성되지 않은 종자(empty seed) 또는 인위적 열처리(95°C)를 가한 종자(dead-filled seed)와 구별시키는 영역대이며, 이는 발아종자가 더 많은 함량의 수분을 보유할 가능성이 높기 때문이라고 설명한 바 있다. 이러한 결과를 고려한다면, Group V의 2개 변수 ab5153(1,941 nm) 및 ab5168(1,935 nm)는 소나무 종자의 수분과 관련되어 있는 것으로 생각된다.

Group VI에는 7개의 변수(XGBoost: 3개, Boosted Tree: 5개)가 포함되어 있으며, 두 모델의 공통변수는 ab4150 (2,409 nm)으로 모델에 대한 총효과는 각각 0.1326 및 0.1597로 매우 높았다. XGBoost 모델의 2개 변수(ab4859 및 ab4844)의 영역대(2,058 nm, 2,064 nm)는 단백질 amide 그룹에서 N-H 신축진동 영역대인 2,055 ~ 2,060 nm(Workman and Weyer, 2012)와 그리고 Boosted Tree 모델의 4개 변수(ab4751, ab4474, ab4196, ab4181)의 영역대(2,104 ~ 2,392 nm)는 목재에서의 셀룰로스 및 리그닌의 O-H 또는 C-H 신축진동 영역대 2,092 ~ 2,384 nm(Schwanninger et al., 2011)와 중첩되어 있었다. 한편, 두 모델 공통변수 ab4150 (2,409 nm)의 영역대는 벤젠 aryl 그룹에서 C-H의 수축 및 굽힘(bending) 진동의 결합 영역대인 2,407 nm와 매우 근접해 있었다. 이상의 결과를 종합하면, Group VI에 속한 7개 변수들의 3개 파장 영역대는 각각 단백질, 셀룰로스 및 리그닌, 그리고 벤젠의 aryl 그룹을 작용기로 갖고 있는 polyunsaturated fatty acids와 관련이 있는 것으로 생각된다.

결론

본 연구는 20년간 장기 저장된 소나무 종자를 FT NIR 스펙트럼으로 측정하여 종자의 발아 유무를 정확히 예측할 수 있는 머신러닝 모델을 확립하기 위해 수행되었다. Savitzky-Golay 2차 다항식 변환을 통해 얻어진 스펙트럼 데이터를 7개의 머신러닝 방법에 적용하여 종자발아 유무 판별을 위한 이항분류 모델을 만들고 예측성능을 비교한 결과, XGBoost(28개 변수) 및 Boosted Tree(33개 변수) 축소모델이 정확도, 오분류율 및 AUC에서 다른 머신러닝 모델보다 더 우수한 성능을 보였다. XGBoost와 Boosted

Tree 축소모형을 구성하고 있는 54개 변수들의 파장 영역대는 크게 6개 그룹으로 나눌 수 있었으며, 이들은 방향족 아미노산, 셀룰로스, 리그닌, 전분, 지방산 및 수분 등과 관련된 것으로 추정되었다.

본 연구에서 확립된 FT NIR 분광 데이터과 머신러닝 방법을 이용한 발아유무 예측 모델은 기존의 발아실험 또는 TTC 검사 방법의 단점인 보존 자원의 소모와 일정 기간의 검사시간 소요 문제를 해결, 보완할 수 있다는 점에서 그 의미를 지닌다고 하겠다. 또한, 본 연구에서 확립된 머신러닝 모델을 구성하는 변수들의 파장 영역대가 종자 발아 및 활력에 영향을 주는 것으로 알려진 지방산 등의 화합물과 수분에 관련된 것으로 추정되었기 때문에 저장 기간에 따른 이들의 정량적, 정성적 변화가 실제 종자활력과 관계가 있는 지를 확인하기 위한 후속연구가 필요하며, 이때 본 연구에서 얻어진 결과는 매우 유용한 기초 자료로 활용될 수 있을 것으로 생각된다.

References

- Chen, Q., Lin, H. and Zhao, J. 2021. Advanced nondestructive detection technologies in food. Springer Nature Singapore. Gateway East, Singapore. pp. 333.
- Chen, T. and Guestrin, C. 2016. XGBoost: a scalable tree boosting system. pp. 785-794. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD '16). Association for Computing Machinery. New York, U.S.A.
- Daneshvar, A., Tigabu, M., Karimidoost, A. and Odén, P.C. 2015. Single seed near infrared spectroscopy discriminates viable and non-viable seeds of *Juniperus polycarpus*. *Silva Fennica* 49(5): 1-14.
- Food and Agriculture Organization of the United Nations. 2014. Genebank standards for plant genetic resources for food and agriculture. FAO Working Group. Rome, Italy. pp. 181.
- JMP Statistical Discovery LLC. 2022. JMP® 17 Documentation Library. JMP Statistical Discovery LLC. North Carolina, U.S.A.
- Kim, D.H., Han, S.H., Song, J.H. and Jang, K.H. 2012. Seed storage and longevity in woody plant. Korea Forest Research Institute, Seoul, Republic of Korea. pp. 159.
- Kim, J.H., Ku, J.J., Lim, H.I. and Kim, Y.Y. 2020. Characteristics and conservation status of useful forest genetic resources seeds. National Institute of Forest Science, Seoul, Republic of Korea. pp. 178.
- Kumari, R. and Srivastava, S. 2017. Machine learning: a review on binary classification. *International Journal of Computer Applications* 160(7): 11-15.
- Lestander, T. and Odén, P.C. 2002. Separation of viable and nonviable filled scots pine seeds by differentiating between drying rates using single seed near infrared transmittance spectroscopy. *Seed Science and Technology* 30(2): 383-392.
- Liu, W., Liu, J., Jiang, J. and Li, Y. 2021. Comparison of partial least squares-discriminant analysis, support vector machines and deep neural networks for spectrometric classification of seed vigour in a broad range of tree species. *Journal of Near Infrared Spectroscopy* 29(1): 33-41.
- Mo, L., Chen, H., Chen, W., Feng, Q. and Xu, L. 2020. Study on evolution methods for the optimization of machine learning models based on FT-NIR spectroscopy. *Infrared Physics and Technology* 108: 103366.
- Mukasa, P., Cho, B.K., Joo, H.J. and Kwon, Y.R. 2018. Determination of viability of Japanese larch seeds using hyperspectral imaging. *Proceedings of the Korean Society for Agricultural Machinery Conference* 23(1): 195.
- Mukasa, P., Wakholi, C., Mo, C.Y., Oh, M.R., Joo, H.J., Suh, H.K. and Cho, B.K. 2019. Determination of viability of *retinispora (Hinoki cypress)* seeds using FT-NIR spectroscopy. *Infrared Physics and Technology* 98: 62-68.
- Narassiguin, A., Bibimoune, M., Elghazel, H. and Aussem, A. 2016. An extensive empirical comparison of ensemble learning methods for binary classification. *Pattern Analysis and Application* 19(4): 1093-1128.
- Qiu, G., Lü, E., Lu, H., Xu, S., Zeng, F. and Shui, Q. 2018. Single-kernel FT-NIR spectroscopy for detecting supersweet corn (*Zea mays* L. *Saccharata* Sturt) seed viability with multivariate data analysis. *Sensors* 18(4): 1010.
- Rocha, W.F.D.C., Prado, C.B.D. and Blonder, N. 2020. Comparison of chemometric problems in food analysis using non-linear methods. *Molecules* 25(13): 3025. <https://doi.org/10.3390/molecules25133025>.
- Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K. and Narasimhan, G. 2020. So you think you can PLS-DA?. *BMC Bioinformatics* 21(Suppl 1): 2.
- Sampaio, P.S. and Brites, C.M. 2021. Near-Infrared spectroscopy and machine learning: analysis and classification methods of rice. pp. 257-288. In: Huang, M. (Ed.). *Integrative Advances in Rice Research*. IntechOpen. London, UK.
- Sato, T., Kawano, S. and Iwamoto, M. 1991. Near-infrared spectral patterns of fatty acid analysis from fats and oils.

- Journal of the Americal Oil Chemists Society 68(11): 827-833.
- Savitzky, A. and Golay, M.J.E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8): 1627-1639.
- Schwanninger, M., Rodrigues, J.C. and Fackler, K. 2011. A review of band assignments in near infrared spectra of wood and wood components. *Journal of Near Infrared Spectroscopy* 19(5): 287-308.
- Shenk, J.S., Workman, J.J. and Westerhaus, M.O. 2007. Application of NIR spectroscopy to agricultural products. pp. 419-474. In: Burns D.A. and Ciurczak E.W. (eds), *Handbook of Near-Infrared Spectroscopy*. CRC Press. New York, U.S.A.
- Shetty, N, Min T.G., Gislum, R., Olesen M.H. and Boelt B. 2011. Optimal sample size for predicting viability of cabbage and radish seeds based on near infrared spectra of single seeds. *Journal of Near Infrared Spectroscopy* 19(6): 451-461.
- Tian, W., Zang, L., Nie, L., Li, L., Zhong, L., Guo, X., Huang, S. and Zang, H. 2023. Structural analysis and classification of low-molecular-weight hyaluronic acid by near-infrared spectroscopy: a comparison between traditional machine learning and deep learning. *Molecules* 28(2): 809.
- Tigabu, M. 2003. Characterization of forest tree seed quality with near infrared spectroscopy and multivariate analysis. (Dissertation). Umeå, Sweden. *Acta Universitatis Agriculturae Sueciae*.
- Tigabu, M. and Odén, P.C. 2003. Discrimination of viable and empty seeds of *Pinus patula* Schiede & Deppe with near-infrared spectroscopy. *New Forest*. 25(3): 163-176.
- Tigabu, M., Daneshvar, A., Jingjing, R., Wu, P., Ma, X. and Odén, P.C. 2019. Multivariate discriminant analysis of single seed near infrared spectra for sorting dead-filled and viable seeds of three pine Species: Does one model fit all species? *Forests* 10(6): 469-482.
- Tigabu, M., Daneshvar, A., Wu, P., Ma, X. and Odén, P.C. 2020. Rapid and non-destructive evaluation of seed quality of Chinese fir by near infrared spectroscopy and multivariate discriminant analysis. *New Forests* 51(3): 395-408.
- Wang, L, Huang, Z. and Wang, R. 2021. Discrimination of cracked soybean seeds by near-infrared spectroscopy and random variable selection. *Infrared Physics and Technology* 115: 103731.
- Workman, J. and Weyer, L. 2012. *Practical guide and spectral atlas for interpretative near-infrared spectroscopy*. 2nd Edition. CRC Press. New York, U.S.A. pp. 326.
- Xia, Y., Xu, Y., Li, J., Zhang, C. and Fan, S. 2019. Recent advances in emerging techniques for non-destructive detection of seed viability: A review. *Artificial Intelligence in Agriculture* 1: 35-47.

Manuscript Received : March 3, 2023

First Revision : April 12, 2023

Accepted : April 13, 2023