The Pharma Innovation Journal 2022; 11(7): 1361-1366

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2022; 11(7): 1361-1366

www.thepharmajournal.com
Received: 02-05-2022
Accepted: 19-06-2022

N Vinoda
Assistant Professor, Department of Processing and Food Engineering, Dr. NTR College of Agricultural Engineering, Bapatla, Andhra Pradesh, India

Premkumar Borugadda
Ph.D. Scholar, Department of Computer Science, Pondicherry University, Karaikal, Puducherry, India

Vimala Beera
Assistant Professor, Department of Food Safety and Quality Assurance, Dr N.T.R College of Food Science & Technology, Bapatla, Andhra Pradesh, India

Ravi Babu M
Assistant professor, Department of plant physiology, Agricultural College, Bapatla, Andhra Pradesh, India

Corresponding Author:
N Vinoda
Assistant Professor, Department of Processing and Food Engineering, Dr. NTR College of Agricultural Engineering, Bapatla, Andhra Pradesh, India

Dimensionality reduction-based approach to classify the cotton leaf images using transfer learning on VGG16

N Vinoda, Premkumar Borugadda, Vimala Beera and Ravi Babu M

Abstract
In Precession agriculture, computer vision has been demonstrated as state-of-the-art technology. In this paper, a VGG16 model was applied to identify and classify cotton leaf diseases. Cotton Dataset consists of 2204 images, in which 1951 images were used for training and 253 images were used for validation. Apply the transfer learning on thirteen convolutional layers of VGG16 for extracting the features on 1951 images. 25088 features are extracted by transfer learning. With these features form high dimensions, if we apply any classification algorithms on high dimension model, may get over fitted. So, for reducing large dimensions, use one dimension reduction technique, namely Principal component analysis (PCA). The output of PCA is low dimension. Now apply three fully connected layers of VGG16 and machine learning classification algorithms on low dimension data. Three fully connected layers of VGG16 provided the best performance model with a 95.65% validation accuracy at the training time of about 140 seconds.

Keywords: Cotton disease detection, machine learning, VGG16, PCA, validation accuracy

Introduction
Computer vision has become a novel technology in various fields of applications such as medicine machine vision. Computer vision performs image capturing, imaging processing, image analyzing, image classification, image reorganization, and named a few advancements in deep learning techniques that have led to automating the many computer vision tasks [1]. Cotton [2] is one of the world's foremost and economy-driven crops for all agricultural-based countries. The reduction in cotton yield led to high economic loss to the farmers. Smart farming is vital to conduct disease incidence at a low level, good management strategies and taking preventive measures at the right time to reduce chemical usage and to increase production. Monitoring the crop during all stages of plant growth requires expert knowledge in the domain and extensive laborious work. Among all deep learning techniques, convolution neural networks are a commonly employed method in image-based data applications. Convolutional Neural Network (CNN) [3] offers feature extraction significantly easier with minimal human supervision and field knowledge than machine learning algorithms. The effectiveness of machine learning algorithms highly depends on the integrity of the input data representation. If the construction of features from raw data is poor, the machine learning algorithms may provide incorrect discrimination between data classes.
Hence, the present study was undertaken to predict the optimal model from various models, namely VGG16 [4] and machine learning models, to classify cotton diseases based on the extracted image features.

Material and Methods
This section discusses data sets and hardware configuration details.

Hardware and Software Specifications
All experiments are performed on a powerful machine, having the specifications are Memory (RAM) 16GB, Processor Intel(R) Core (TM) i7-10875H CPU @ 2.30GHz 2.30 GHz, Graphics (GPU) NVIDIA GeForce RTX 2070- 8GB, Operating system Windows 10, 64 bits, Integrated Development Environment (IDE) Jupyter Notebook.

Methodology

The Research framework has four phases and is shown in figure 1. The first phase has to pre-process. The second phase has a feature extraction process that includes the CNN training techniques, namely VGG16. Dimension reduction is the third phase. The fourth phase had classification algorithms.

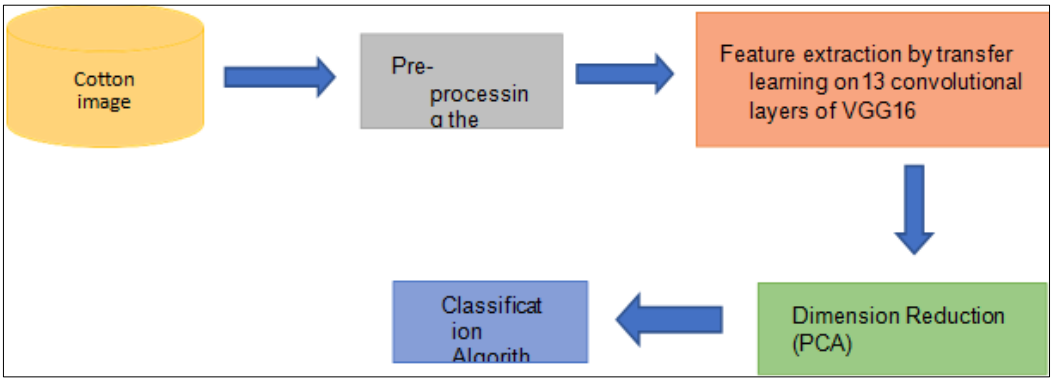

Fig 1: Framework for Disease Classification

Datasets

An open-access cotton disease dataset [5] is used for training and validating the VGG16 model. The number of images in each training set and validating set contain 1951 and 253, respectively. The detailed information about the dataset, such as training and validation, is given in Tables 2 and 3. This dataset contains four distinct categories of images: fresh cotton leaves, fresh cotton plants, diseases cotton leaves, and diseased cotton plants, which are imbalanced datasets. Four classes in the training dataset don't have an approximately equal proportion.

Table 2: Cotton train dataset

Type of Dataset	Category	No. of Images	Percentage of classes (%)
	Diseased cotton leaf	288	14.76
Train data	Diseased cotton plant	815	41.77
	Fresh cotton leaf	427	21.88
	Fresh cotton plant	421	21.57
	Total No. of training images	1951	

Table 3: Cotton validation dataset

Type of Dataset	Category	No. of Images	Percentage of (%) classes
Validation data	Diseased cotton leaf	43	16.99
	Diseased cotton plant	78	30.88
	Fresh cotton leaf	66	26.08
	Fresh cotton plant	66	26.08

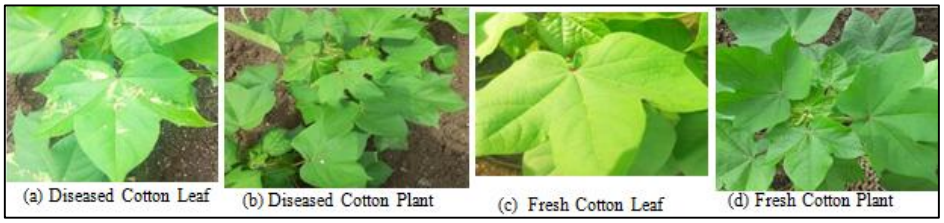

Fig 2: Sample images of cotton leaves and plants

Preprocessing Phase

The first phase is preprocessing of input image data with a size (h, w, c) of 227,227,3 that has been done in a sequence of operations. The input dataset classes are labelled through label encoding then apply a one- hot encoding technique. Here, class labels, namely disease cotton leaf, disease cotton plant, fresh cotton leaf and the fresh cotton plant, are text data. The system can't understand the text data. So, we need to convert this kind of categorical text data into model-understandable numerical data with label encoding. The Lael encoding method will assign numbers between 0 and n-1, where n is a number of class labels (n=4) based on alphabetical order. Besides, one hot Encoding is another technique to treat categorical variables. This creates additional attributes based on the unique value in the categorical variable [6, 7]. Then the pixel values of images are normalized between 0 and 1.

Feature Extraction Phase

In this feature extraction phase, the standard VGG16 model has been applied to extract features. Apply transfer learning [8] on 13 Convolutional layers of VGG16 to extract the number of features is shown in figure 3.

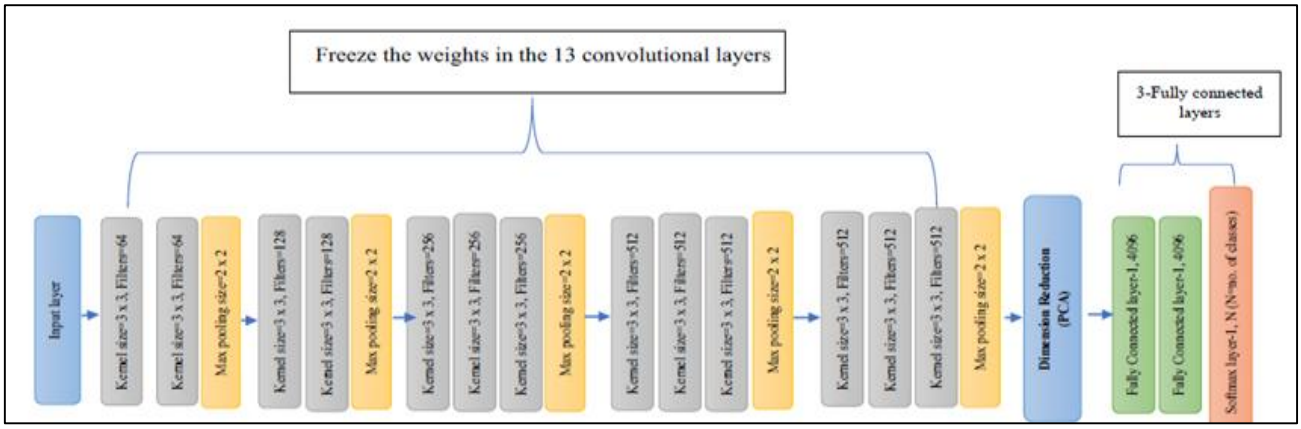

Fig 3: Applying dimension reduction methods between transfer learning on 13 Convolutional layers of VGG16 and 3 fully connected layers

Dimension Reduction Phase

In the previous stage, high dimensional 25088 feature space has been obtained with 13- convolutional layers and transfer learning on VGG16. If we train the classification models like machine learning algorithms (MLA) and three fully connected layers of VGG16 with high dimensional features, models will face some issues like, there is a chance that the model will be biased towards over fitting, model computation will be high, the performance of models will be low and curse of dimensionality. So, to address all these issues, we need to apply dimension reduction like PCA [9, 10] to reduce the dimension. PCA with 99% of the variance and formed as a new set of components are 1704 shown in figure 4.

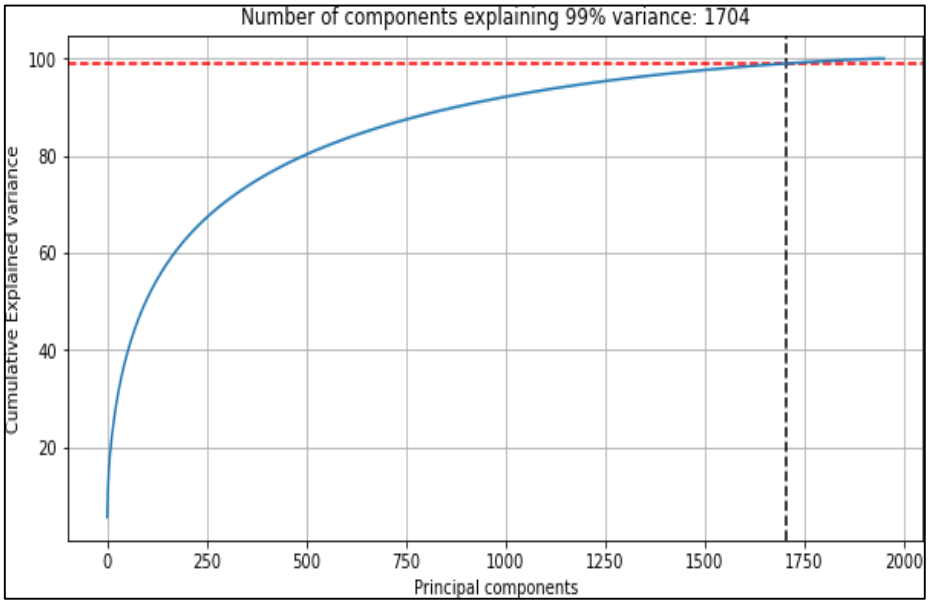

Fig 4: Number of components explained 99% of variance are 1704

Classification phase

Apply the traditional machine learning algorithms, namely, ABC [11, 12], DTC [13, 14], GBC [15, 16], KNN [17, 18], LR [19, 20], RFC [21, 22], SVC [23, 24] and 3fully connected layers of VGG16 to lower dimension components 1216 in classification phase. While training VGG16 model and MLA apply the hyper parameters for better results are shown in table 4 and 5.

Table 4: Hyper parameters of Deep learning

S. No	Hyper parameter	Values
1	Activation functions	Relu, softmax
2	Optimizers	SGD, Adam
3	Learning rate	0.1,0.001,0.0001,0.00001
4	Dropout	0.2,0.3,0.4,0.5
5	Decay	1e-3, 1e-4, 1e-5, 1e-6
6	Momentum	0.8,0.9
7	Patience	15,20,30
8	Minimum delta	0.01, 0.001, 0.0001
9	Batch size	8,16,32,64,128,256
10	Epochs	100, 500, 1000

Table 5: Hyper parameters of ML Algorithms

S. No	Machine Learning Algorithm	Hyperparameter	Values
1	ABC	Learning rate	0.01
		n_ estimators	200
		Criterion	entropy
2	DTC	max_ depth	5
		min_ samples	2
		min_ samples_ split	2
		Algorithm	Ball_ tree
		leaf_ size	20
3	KNN	Metric	minkowski
		n_ neighbors	5
		P	2
		Weights	distance
4	LDA	solver	svd
		C	0.01
5	LR	max_ iter	100
		Penalty	l2
		solver	liblinear
		Criterion	gini
		Max_ depth	80
6	RFC	max_ features	0.33
		min_ samples_ leaf	2
		min_ samples_ splt	2
		n_ estimators	300
		C	10
7	SVC	Gamma	0.0001
		Kernel	rbf
8	XGB	Eta	0.01
		Max_ depth	5
		gamma	0

Results and Discussion

Performance Measure

Evaluated the performance of classification models through a confusion matrix from figure5. Evaluation metrics are Accuracy, Precision, Recall (Sensitivity), Specificity, F1_Score. F1 score, which is used when the dataset belongs imbalanced. The given dataset is imbalanced so, we need to choose the optimal model based on macro average f1_score. True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) parameters of the confusion matrix were used to calculate the metrics [25, 26] for 'k' classes.

	Predicted Labels					
ACTUAL LABELS		L1	L2	…	LK	
	L1	TP1				TPR1
	L2		TP2			TPR2
	L3	…	…	…	…	…
	LK				TPK	TPRK
		PPV1	PPC2	…	PPVK	

Fig 5: The confusion matrix for the multiclass classification

Experimental Results

VGG16 Results

Three fully connected layers of VGG16 and machine learning algorithms are applied to 1704 principal components in the classification phase, and results are shown in below table 6 and 7. The best results are 95.65% of a validation score, and 95.19%macro average of F1_score found from table 6 at hyperparameters batch size 128, learning rate 0.0001, decay value is 1e-6, momentum is 0.9 and number of epochs are 275 for getting these optimal results to train the VGG16 model for approximately 140 seconds.

Table 6: Results of VGG16 on 1704 principal components

B. S	E	Train time	T.A (%)	T. L	V.A (%)	V. L	M.A.P (%)	M.A.R (%)	M.A.F1 (%)	Storage Space
		(H:M:S)								
8	103	0:22:12	100.0	0.0007	94.47	0.14	94.78	93.77	94.16	181 MB
16	135	0:03:04	100.0	0.0015	95.65	0.14	95.83	95.00	95.34	181 MB
32	196	0:03:23	100.0	0.0014	94.47	0.14	94.78	93.77	94.16	181 MB
64	210	0:02:48	100.0	0.0025	93.68	0.15	93.80	93.02	93.34	181 MB
128	275	0:02:20	100.0	0.0048	95.65	0.12	95.52	94.94	95.19	181 MB

256	340	0:01:58	100.0	0.014	94.86	0.16	95.13	94.50	94.76	181 MB

B.S-Batch size; E-Epochs, T.A-Train Accuracy; T.L-Train Loss; V.A-Validation Accuracy; V.L-Validation Loss; M.A.P-Macro Average Precession, M.A.R-Macro Average Recall; M.A.F1-Macro Average F1_Score.

Table 7: Results of classification models on 1704 principal components

Model	Train Time (H:M:S)	T. A (%)	V. A (%)	M.A.P (%)	M.A.R (%)	M.A.F1 (%)	Storage Space
LR	0:00:02	99.90	94.86	95.11	94.61	94.80	54.1 KB
RFC	0:02:54	100.0	85.38	87.16	83.50	84.53	6.74 MB
DTC	0:00:03	85.90	77.08	77.23	77.08	76.79	11.0 KB
ABC	0:00:58	69.04	64.43	72.65	62.23	59.13	140 KB
KNN	0:00:01	100.0	71.14	80.29	66.33	65.73	39.7 MB
SVC	0:00:05	99.89	94.07	94.81	93.39	93.95	14.7 MB
XGB	0:00:20	100.0	86.96	88.12	85.62	86.39	750 KB
LDA	0:00:01	100.0	92.89	93.46	92.84	93.10	160 KB

The comparison of M.A.P., M.A.R., M.A.F1 & V.A. at different batch sizes are shown in table 6 at batch size 128, the best validation accuracy and M.A.F1_ score is 95.65% and 95.19%, respectively. Different machine learning classification model results are shown in table 7. Among eight models, the logistic regression (L.R.) model has the highest validation score, 94.86% and 94.80% of M.A.F1_score. Training and validation losses training and validation accuracies are shown in figure 8 and figure 9, respectively. The confusion matrix and classification report of VGG16 are shown in Figures 10 and 11, respectively.

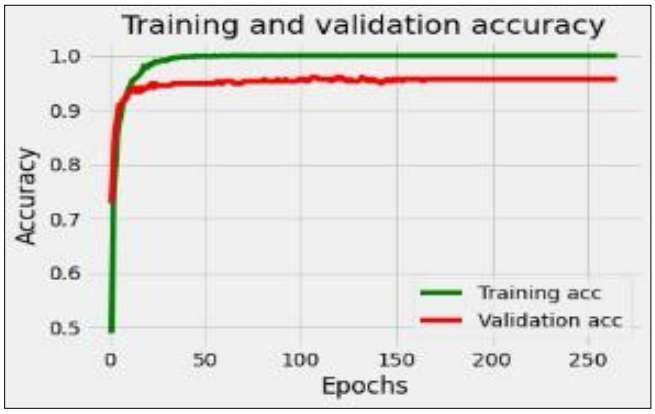

Fig 6: Train & validation accuracy of VGG16 at batch size 128

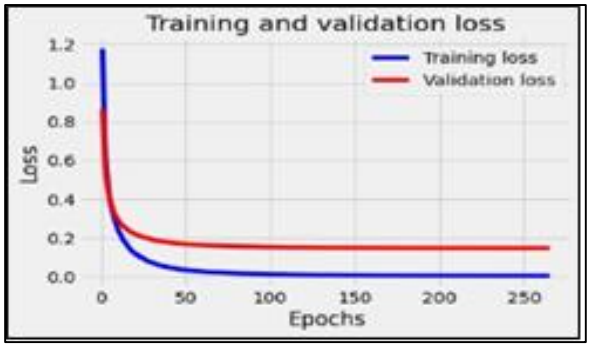

Fig 7: Train & validation loss of VGG16 at batch size 128

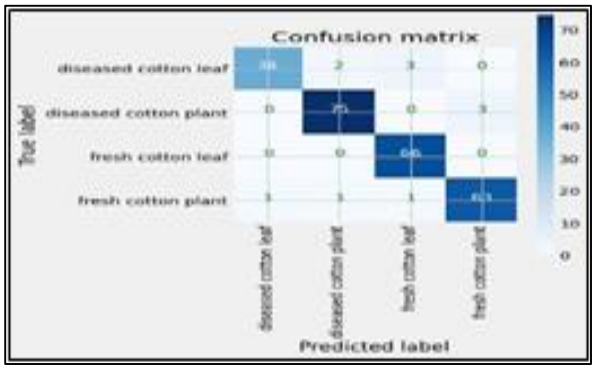

Fig 8: Confusion matrix of VGG16 at batch size 128

Classification report :	precision	recall	f1-score	support
diseased cotton leaf	0.9744	0.8837	0.9268	43
diseased cotton plant	0.9615	0.9615	0.9615	78
fresh cotton leaf	0.9429	1.0000	0.9706	66
fresh cotton plant	0.9545	0.9545	0.9545	66
accuracy			0.9565	253
macro avg	0.9583	0.9500	0.9534	253
weighted avg	0.9570	0.9565	0.9562	253

Fig 9: Classification Report of VGG16 at batch size 128

Conclusion

Thirteen convolutional layers of the VGG16 model with transfer learning is used to extract the features from images, and those features are fed to PCA for dimension reduction. The output of PCA is 1704 principal components are fed as input to three fully connected layers of VGG16 and machine learning classification models for classifying the cotton leaf disease. Among all models, the VGG16 model has given the best results. Given dataset is imbalanced data, for imbalanced data, based on the macro F1_score, we choose the optimal

deployment model. Among all classification models, the VGG16 model has given the best result of F1_score is 95.34%.

5. References

1. Patrício DI, Rieder R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. Computers and electronics in agriculture. 2018;153:69-81.
2. John ME. Cotton crop improvement through genetic engineering. Critical Reviews in Biotechnology. 1997;17(3):185-208.
3. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET) IEEE. 2017, 1-6.
4. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint ar. 2014;Xiv:1409.1556.
5. https://www.kaggle.com/janmejaybhoi/cotton-disease-dataset
6. Gu B, Sung Y. Enhanced reinforcement learning method combining one-hot encoding- based vectors for cnn-based alternative high-level decisions applied sciences. 2021;11(3):1291.
7. Yang X, Hou L, Zhou Y, Wang W, Yan J. dense label encoding for boundary discontinuity free rotation detection in proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021;15819-15829.
8. Tammina S. Transfer learning using VGG-16 with deep convolutional neural network for classifying images. International Journal of Scientific and Research Publications (IJSRP). 2019;9(10):143-150.
9. Mudrova M, Procházka A. Principal component analysis in image processing. In Proceedings of the MATLAB technical computing conference, Prague, 2005.
10. Vidal R, Ma Y, Sastry S. Generalized principal component analysis (GPCA). IEEE transactions on pattern analysis and machine intelligence. 2005;27(12):1945-1959.
11. Washburn PS. Investigation of severity level of diabetic retinopathy using adaboost classifier algorithm materials today: proceedings. 2020;33:3037-3042.
12. Kumar CS, Sharma VK, Yadav AK, Singh A. perception of plant diseases in color images through Adaboost in innovations in computational intelligence and computer vision. Springer, Singapore. 2021, 506-511.
13. Priyanka, Kumar D. decision tree classifier: a detailed survey. International journal of information and decision sciences. 2020;2(3):246-269.
14. Koga S, Zhou X, Dickson DW. Machine learning-based decision tree classifier for the diagnosis of progressive supranuclear palsy and Corticobasal degeneration. Neuropathology and applied neurobiology. 2021.
15. Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. boosting methods for multiclass imbalanced data classification: an experimental review. Journal of big data. 2020;7(1):1-47.
16. Shrivastav LK, Jha SK. a gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of covid-19 in India applied intelligence. 2021;51(5):2727-2739.
17. Abu Alfeilat HA, Hassanat AB, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big data. 2019;7(4):221-248.
18. Patil A, Lad K. chili plant leaf disease detection using SVM and KNN classification in rising threats in expert applications and solutions. Springer, Singapore. 2021, 223-231.
19. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis international journal of medical informatics, 2021, 104484.
20. Xiao R, Cui X, Giao H, Zheng X, Zhang Y. early diagnosis model of Alzheimer's disease based on sparse logistic regression. Multimedia tools and applications. 2021;80(3):3969-3980.
21. More AS, Rana DP. Review of random forest classification techniques to resolve data imbalance in 2017 1st international conference on intelligent systems and information management (ICISIM). 2017, October, 72-78. IEEE.
22. Saha S, Ahsan SMM. Rice disease detection using intensity moments and random forest in 2021 international conference on information and communication technology for sustainable development (icict4sd) IEEE. 2021, 166-170.
23. Ghaddar B, Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. European journal of operational research. 2018;265(3):993-1004.
24. Raghavendra Y. Multivariant disease detection from different plant leaves and classification using multiclass support vector machine. Turkish journal of computer and mathematics education (Turcomat). 2021;12(13):546-556.
25. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information processing & management. 2009;45(4):427-437.
26. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: An overview. arXiv preprint. 2020;arXiv:2008.05756.