Fungal Blast Disease Detection in Rice Seed using Machine Learning

Raj Kumar[1], Gulsher Baloch[2], Pankaj[3], Abdul Baseer Buriro[4], Junaid Bhatti[5]
Department of Electrical Engineering, Sukkur IBA University, Sukkur, Pakistan

***Abstract*—The economy of Pakistan mainly relies upon agriculture alongside other vital industries. Fungal blast is one of the significant plant diseases found in rice crops, leading to reduction of agricultural products and hindrance in the country's economic development. Plant disease detection is an initial step towards improving the yield and quality of agricultural products. Manual Analyzation of plant health is tiresome, time taking and costly. Machine learning offers an alternate inspection method providing benefits of automated inspection, ease of availability, and cost reduction. The visual patterns on the rice plants are processed using the machine learning classifiers such as support vector machine (SVM), logistic regression, decision tree, Naïve Bayes, random forest, linear discriminant analysis (LDA), principal component analysis (PCA), and based on classification results plants are recognized as healthy or unhealthy. For this process, a dataset containing 1000 images of rice seed crop is collected from different fields of Kashmore, and whole analysis of image acquisition, pre-processing, and feature extraction is done on the rice seed only. The dataset is annotated with healthy and unhealthy samples with the help of a plant disease expert. The algorithms used for processing data are evaluated in terms of F1-score and testing accuracy. This paper contains results from traditional classifiers, and alongside these classifiers, transfer learning has been used to compare the results. Finally, a comparative analysis is done between the results of traditional classifiers and deep learning networks.**

Keywords—Fungal blast; machine learning; support vector machine (SVM); logistic regression; decision tree; Naïve Bayes; random forest; linear discriminant analysis (LDA); principal component analysis (PCA); image acquisition; pre-processing; feature extraction; F1-Score; convolutional classifier; deep learning

I. INTRODUCTION

Rice is one of the major agricultural crops in Pakistan, which has a great influence on the country's economy. It is subject to different diseases in its leaves, root, and seed, which may reduce its yield and lead to a reduction in agricultural products [1]. Farmers do not have a specific idea regarding pesticides as per diseases on rice crops [2]. Hence, the rice seed health monitoring with the help of image processing and machine learning algorithms plays an important role in increasing the yield and production of rice [3]. Different related work has been done using machine learning algorithms on rice as well as other crops, which is discussed in Section II. The uniqueness of this research is a dataset of rice seed which is mentioned in Section V. The image processing helps to visualize the plant's images clearly while removing the extra background and extracting the infected region of the plant with the help of feature extraction and segmentation [4]. All the image processing and classification techniques that have been used are mentioned in the proposed workflow in Section III. Machine learning helps to analyze the plant's health based on the extracted features or cropped images of the dataset [4] [5]. With the help of this process, a disease can be detected in rice crops, and based on that disease, farmers can use the specific pesticide, which will lead to a reduction in cost and time [5]. The current methods for rice disease detection in Pakistan involve the experience of farmers in detecting rice disease, which is not very reliable. Further, the inspection by the disease detection expert is too costly, and local farmers are unable to afford it. This, in turn, affects the production and yield of the rice crops. With the recent advancement in machine learning, this paper proposes the vision-based approach to detect rice plant disease. One of the critical requirements of any machine learning problem solution is data generation and collection. Further, for the machine learning technology to be implemented in real-time requires the handling of different image vision problems such as occlusion detection, background/foreground detection, suitable feature selection, and extraction from the rice crop images to complete the required disease detection task. To imitate the real-time solution implementation, the image data of rice plant from a number of different rice fields in Kashmore, city of Pakistan, has been collected, and with the help of a disease detection expert, the data set has been labeled. Further, recent and state-of-the-art machine learning algorithms are implemented and tested on the dataset for rice disease detection, and the results are compared in terms of F-score and accuracy. All the proposed work which successfully have been implemented is mentioned in section IV, which has multiple results. A final conclusion has been made over different classification results, which is mentioned in Section VI. This paper helps summarize the recent and state-of-the-art algorithms for rice disease detection and also helps the authors to cater upon problems for the implementation of the algorithms in real-time rice disease detection.

II. LITERATURE REVIEW

Kawcher Ahmed et al. [7] implemented a machine learning algorithm for the detection of three common rice leaf diseases which are leaf smut, bacterial leaf blight, and brown spot diseases. The dataset used was already refined and collected from an online website [8]. For classification purposes, KNN (K-Nearest Neighbor), Decision Tree, Naïve Bayes, and Logistic Regression [8] [9] are used. It is concluded that the decision tree algorithm after 10-fold cross-validation has better performance with an accuracy of 97% applied on the test dataset. Neha G. Kurale et al. [9] analyzed leaf diseases in plants generally using the texture features and neural network. They summarized that for the plant's leaf disease detection, support vector machine (SVM), KNN (K-Nearest Neighbor), and Neural Networks techniques [9] have the most appropriate and effective results. Anjna et al. [10] have worked on capsicum disease symptoms, and she has used k-means clustering, BPNN classifier, neural network classifier, thresholding-based segmentation, minimum distance criterion, and SVM [10]. The authors have extracted GLCM features on which they have classified the capsicum of diseases. The SVM and KNN classifiers have 100% accuracy being the highest [10]. It is concluded that neural network classifier gives better results as compared to others in a short time with texture, shape, and co-efficient features [11] [12]. Naga Swetha R. et al. [13] analyzed and detected four different diseases in rice plants which are the bacterial blight of rice, rice blast, and false smut. The dataset of total 115 rice disease images have been collected by themselves and some have been collected from the internet [13]. Only two classifiers, support vector machine (SVM) and KNN (K-Nearest Neighbor) are used based on shape and color features [13]. A mobile application for the automatic diagnosis of diseases in rice plants has also been developed [13]. Muhammad Kashif et al. [14] analyzed the different feature techniques regarding plant disease detection generally. The authors used the texture, Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Binary Robust Invariant Scalable Keypoints (BRISK), Binary Robust Independent Elementary Features (BRIEF), and Fast Retina Keypoints (FREAK) features for plant disease detection [14]. They concluded that dense SIFT features give the best results with an accuracy of 98.36% [14] [15]. Harshadkumar B. Parjapati et al. [16] analyzed and implemented a machine learning algorithm for the three different leaf diseases of rice plants which are bacterial leaf blight, brown spot, and leaf smut. They collected datasets from the rice fields [16]. They have applied three different techniques of segmentation and for the accurate features, they used K-means clustering segmentation [16] [17]. For the classification, they used an SVM classifier based on color, shape, and texture features [17]. They got an accuracy of 93.33% for the training dataset and 73.33% for the testing dataset. They have also applied k-fold cross-validation and got an accuracy of 83.80% for 5-folds and 88.57% for 10-folds [17]. Efetkhar Hossain et al. [18] used only KNN (K-Nearest Neighbor) classifier based on texture features for the detection of plant diseases. They used the dataset of 237 plant leaf images that were already refined and have been collected from two different database websites [18]. They proposed that the KNN classifier can classify the diseases like Alternaria alternate, anthracnose, bacterial blight, leaf spot, and canker of various plant species [18]. They concluded that the proposed KNN classifier with texture features could detect diseases with 97.76% accuracy [18]. Budiarianto Suryo Kusumo et al. [19] proposed a machine learning algorithm for disease detection in the Corn crop. The dataset used was already refined and has been collected from the PlantVillage dataset website [19]. They used several image processing techniques for feature extraction such as SIFR, SURF, BRIEF, and HOG [19][20][26]. For classification purposes, they used SVM, decision tree, random forest, and Naïve Bayes algorithms [20]. Finally, it is concluded that the color features are most important for disease detection in the corn crop. Sandeep Kumar et al. [21] used support vector regression (SVR) with different classification based on shape, color, texture, and cosine features of plant species for plant disease detection. The authors used a limited plant leaf dataset that has been collected by themselves. They proposed three different computer vision techniques for plant disease detection which are feature discovery, feature explanation, and image depiction [5] [4] [13]. The proposed approach uses SIFT and SURF features and the clustering is done by F-Dbscan [5]. Sachin D. Khirade et al. [1] discussed the different techniques and processes for plant health monitoring and disease detection. The dataset they used, is captured by themselves [1]. They proposed the image processing techniques such as image pre-processing and image segmentation are the most useful for plant disease detection [4] [6] [7]. They used different feature extraction techniques for the extraction of texture, shape, and color features. For classification purposes, they used ANN (Artificial Neural Network) such as self-organizing feature map, back propagation algorithm, and SVM [12]. Pushkara Sharma et al. [19] conducted a study in India on various plant leaves to detect the diseases using pre-processing techniques and segmentation to get the useful part of the leaf. After preprocessing and segmentation, they used Logistic regression, KNN, SVM, and CNN classifiers [19]. The highest accuracy that he got was 98.0% from the CNN model. The authors proposed that through segmentation, the diseased portion of the input image can be detected [21]. For the feature extraction, different feature extraction techniques and different classifiers are used. Arsa, D. M. S et al. [22] has used VGG-16 pre-trained model in Batik based on random forest. They have used precision, recall, F-score, and accuracy to evaluate their proposed method performance [22]. Ufaq Khan et al. [25] divided plant disease detection techniques into two phases; the first is segmentation, and the other is classification. In this paper, the author generally described the techniques for plant disease detection, so they did not use any dataset [25].

After reviewing all the mentioned studies, the proposed work is novel because in the above studies, mostly plant dataset used consists of less than 300 images from one field, and mostly dataset has been collected from the internet, which was already refined and did not necessarily reflect the real field scenario. But in this case, the unique dataset of 1000 healthy and unhealthy rice seed images have been captured from different rice fields. Another uniqueness from the above studies is that most have extracted limited image features while in this case, three different types of features of an image, such as texture, SURF, and BRISK features have been extracted. Moreover, for the testing and training results in the above studies, limited classifiers have been used, such as SVM and decision tree, while in this case, six different classification algorithms such as SVM, LDA, decision tree, logistic regression, Naïve Bayes, and random forest have been used. For the most accurate results, the dataset has been used with different image sizes such as 128x128, 256x256, 512x512, and 1024x1024. PCA and k-fold cross-validation have been applied to every classifier for better accuracy performances, and finally, the comparatively better results are with SVM and random forest classifiers.

III. Proposed Methodology

In this section, a complete methodology for fungal blast disease detection has been proposed in a block diagram, shown in Fig. 1. Every step is performed for the best accuracy results. In image acquisition, a unique dataset has been collected, and different image processing techniques are applied, such as image cropping, color enhancement, and image resizing for a better understanding of the dataset. Further, feature extraction techniques are used, such as BRISK, SURF, and texture features, to remove the extra background and to get the infected region of dataset. The extracted features are used for the classification purposes while taking 80% of the training dataset and 20% of the testing dataset. Different classifiers such SVM, LDA, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes and PCA classifiers with 10-fold cross validation are used for a comparative analysis based on F1-score and testing accuracy. A descriptive analysis is given as under:

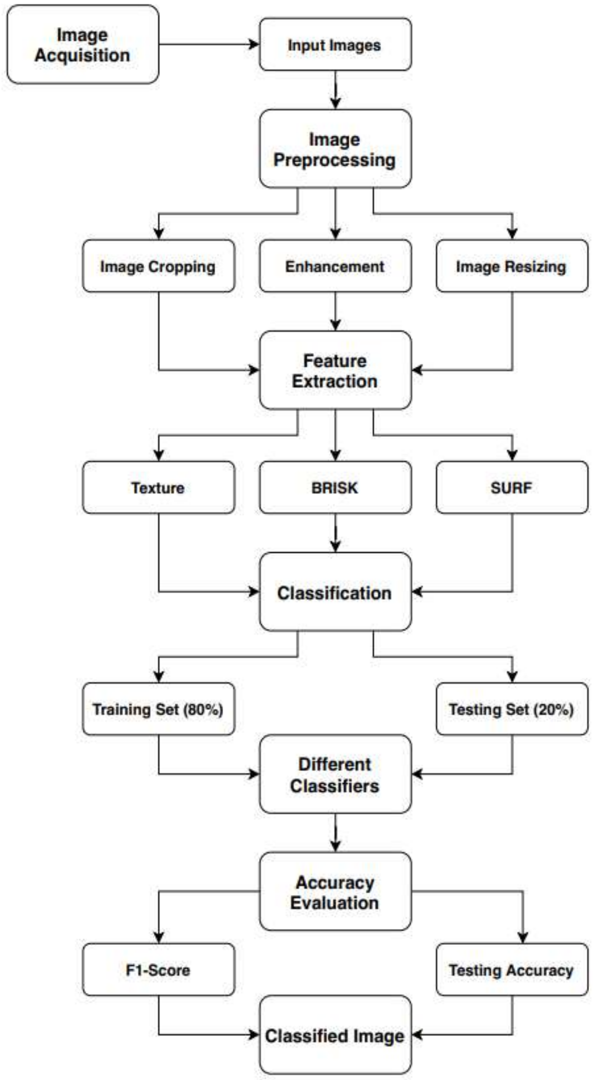
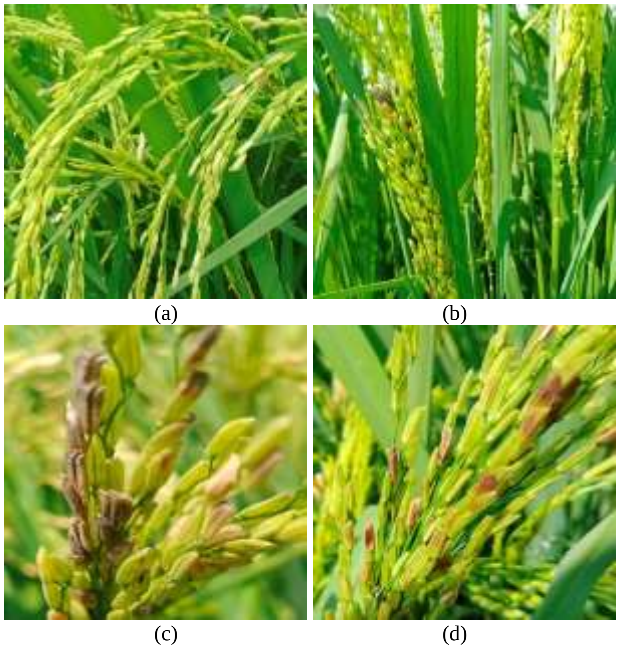


Fig. 1. Proposed Work Flow Chart for Traditional Classification.

A. Image Acquisition

A unique dataset of healthy and unhealthy rice crop has been captured through an android camera from the different fields of Kashmore. The dataset has been captured from September 5th to 7th, 2020 and the age of the crop at that time was 50 to 60 days. The captured images were in RGB (Red, Green, and Blue) form. The whole dataset consists of both healthy and unhealthy crops of 1300 different data samples of rice seed plants annotated with the help of a plant disease expert.

1) Dataset description: Initially, a total of 1500 images of healthy and unhealthy rice crops have been captured from the field. Due to huge distortion in the background and extra parts, images that were not helpful have been removed, and finally, the 1000 healthy and unhealthy images are left in the dataset. The dataset is uploaded on “Kaggle” website, which is now open to use for everyone. The sample images of the healthy and unhealthy dataset are shown in Fig. 2.



(a) (b)

(c) (d)

Fig. 2. Sample Images from the Dataset of Rice Seed Plant uploaded on Kaggle: (a) and (b) are Healthy Plants of Rice Crop While (c) and (d) are unhealhty Crop because it has Brownish Spots on Seeds.

B. Image Pre-Processing

different pre-processing techniques have been used to prepare data for machine learning classification and evaluation, such as image cropping, image resizing, and image enhancement [20] [21] [22].

1) Image cropping: Image cropping has been performed manually for every image to remove the extra part from the images [17] [18] [20].

2) Image resizing: Image resizing has been done to take all the datasets of equal size, which will help in feature extraction to get the balanced features [6] [7] [13]. For the comparison of better results, all the image sizes have been taken, such as 128 x 128, 256 x 256, 512 x 512, and 1024 x

1024. Better results have been achieved for the image size of 256 x 256 in every classifier. The comparison histograms are shown in Fig. 6.

3) Image enhancement: Image enhancement has been performed for the whole dataset to increase the contrast of images. The RGB dataset has been converted into grayscale for better performances [6] [7] [13].

C. *Feature Extraction*

Feature extraction is a key step to analyze the image deeply with the help of features. It helps to get useful information from the image [1]. Multiple features of the rice plant dataset have been extracted, such as Gray level co-occurrence matrix (GLCM) or texture features [1] [3], brisk and surf features [14] [3], shown in Table I. This table shows the feature types and their name that have been extracted from the dataset of rice crops. A total of three feature types have been extracted, and normalization is applied for all the features before classification.

1) Texture features: Texture features define the distribution of color, roughness, and hardness in an image. It helps mainly for the detection of infected areas in the image of rice crop [5]. Texture-based features are contrast, correlation, energy, entropy, and homogeneity [14]. Contrast is the intensity measurement between a pixel and its neighbor in an image. Correlation defines that how correlated a pixel is with its neighboring pixel in the entire image. Energy is the measurement of uniformity which means how much homogeneous an image is, the large the energy. Entropy is the measurement of image intensity or disorder. Homogeneity defines the similarity of pixels in an image [13]. Equations for all the texture features or gray level co-occurrence matrix of these features are shown in Table II.

2) Speeded-Up Robust Features (SURF): The SURF algorithm is related to the Scale Invariant Feature Transform (SIFT). It is used to detect the local features of an image in a very quick and reliable manner [10]. In SURF, first of all, the key-points of an image are perceived, and then related consistent descriptors are calculated [6].

3) Binary Robust Invariant Scalable Keypoints (BRISK): BRISK is a binary descriptor in which key-points are selected, and then a sampling pattern is applied to the neighbors of those key-points in an image. Every pair of pixels around the key-points is separated by two subsets, such as long-distance pair and short-distance pair [14].

D. *Classification*

Classification is important for the detection of fungal blast disease in rice crops. It imposes a class on the new sample with the help of learning from different classifier models by training [3]. Classification can be performed by using the actual image of the dataset or by using the features which have extracted. The main reason purpose of using classification is, it can detect plant disease automatically [9]. Classification with traditional classifiers can be done with the help of features. For the classification of rice crops, both convolutional and traditional classifiers have been used. All the feature values have been given as input to the below-mentioned classifiers by splitting 80% of data for training and 20% for testing.

1) Support Vector Machine (SVM): SVM is a supervised learning algorithm that uses Support Vector Classification (SVC) for classification purposes. It is a linear classification technique and has been found most competitive in machine learning algorithms for the classification of high-dimensional datasets [10]. SVM is easy to use and controls the complexity of decision and frequency error [20]. Equation (6) shows how the SVM classifier works at the backend. The accuracy achieved in the SVM classifier with the image size of 256 x 256 has the highest accuracy before PCA [6] as compared to other classifiers. The accuracy comparison of SVM with different sizes of the dataset is shown in Fig. 6.

$$argmax_{j=1\ldots M}\ g^{j}(x)\ where\ g^{j}(x)=\sum_{\mathrm{i}=1}^{\mathrm{m}} y_i a_i^j k(x,x_i)+b^j \quad (6)$$

2) Linear Discriminant Analysis (LDA): LDA is a supervised learning algorithm that finds the linear combination based on different features that can split two or more classes. It can also be used for dimensionality reduction purposes because it can be used for more than two classes for classification [21]. Like SVM, it is a linear classification technique. Equation (7) shows the discriminant for the linear variable, so this is the equation for the linear discriminant. The accuracy achieved in the LDA classifier before PCA for 256 x 256 image size is comparatively less than SVM classifier. The accuracy comparison for different image sizes is shown in Fig. 6.

$$\delta_k(k)=x\frac{\mu_k}{\sigma^2}-\frac{\mu_k^2}{2\sigma^2}+\log(\pi_k) \quad (7)$$

TABLE I. FEATURES THAT HAVE BEEN USED

Sr. No.	***Features Type***	***Features Name***
1	Texture Features	Contrast, Correlation, Energy, Entropy, Homogeneity
2	Brisk Features	Scale, Orientation, Metric
3	Surf Features	Scale, Orientation, Metric

TABLE II. FORMULA FOR TEXTURE FEATURES

Eq. No.	***Features Type***	***Features Formula***
1	Contrast	$\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{j=1}^{n}(i,j)^2 p(i,j)$
2	Correlation	$\sum_{\mathrm{i,j}=1}^{\mathrm{n}}\frac{p_{i,j}(i-\mu)(j-\mu)}{\sigma^2}$
3	Energy	$\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{j=1}^{n}\left(p(i,j)\right)^2$
4	Entropy	$\sum_{\mathrm{k}=0}^{\mathrm{i}=1} prk(log_2 prk)$
5	Homogeneity	$\sum_{\mathrm{i,j}=1}^{\mathrm{n}}\frac{p_{i,j}}{1+(i-j)^2}$

3) Logistic Regression (LR): Logistic regression is a statistical supervised machine learning algorithm that is used for classification purposes. It works based on the concept of probability, so it is also known as a predictive analysis algorithm. It uses the complex cost function known as 'Sigmoid function' instead of a linear cost function that is why sometimes it is not said as linear regression [23]. Equation (7) shows the complex cost function of logistic regression, and equation (8) is used for the multiple regression problems, which take more than one predictor. The results for multiple logistic are comparatively better than linear regression. The accuracy achieved in the LR classifier before PCA for an image size of 256 x 256 is smaller than both SVM and LDA classifiers. The accuracy comparison histogram for logistic regression classifier for different image sizes is given in Fig. 6.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases} \qquad (8)$$

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (9)$$

4) Naïve Bayes (NB): Naïve Bayes is a probabilistic algorithm that works based on the Bayes' theorem. This classifier takes every feature conditionally independent with others [10]. With this assumption, it calculates the likelihood of the data using Bayes' theorem with the product of conditional probability [24]. The best hypothesis in Naïve Bayes theorem can be chosen based on equation (10). The accuracy achieved in the NB classifier is the lowest accuracy than all other classifiers. The accuracy comparison plots are given in Fig. 6.

$$\hat{y} = argmax\ P(y)\ \prod_{i=1}^{n}\left(P(x_i|y)\right) \qquad (10)$$

5) Decision Tree (DT): The decision tree is the most useful classifier in machine learning algorithms because it takes the most suitable attribute at its root node [23]. It works based on the entropy and information gain approach for the construction of its tree. Equation (11) shows the formula for entropy, and equation (12) is for gain. If the entropy is more positive, then the instances will be more heterogeneous [24]. The accuracy achieved in the DT classifier for 256 x 256 image size is more than SVM and all other classifiers before PCA. The accuracy comparison histograms are given in Fig. 6.

$$E = \sum_{\mathrm{i=1}}^{\mathrm{c}} -p_i log_2 p_i \qquad (11)$$

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \qquad (12)$$

6) Random Forest (FR): Random forest is a supervised machine learning algorithm, mainly used for classification purposes. It has comparatively better accuracy results than that decision tree classifier because it works based on a decision tree [7] [9] [10]. Random forest is mostly used to avoid overfitting in decision tree classifiers. It constructs the trees which have been trained using the data samples training and features [10]. The accuracy achieved in the RF classifier for 256 x 256 image size before PCA is greater than all other classifiers. The accuracy comparison histograms are shown in Fig. 6. A random forest classifier has been concluded best for the fungal blast disease detection based on already defined features. The feature importance graph for random forest classifier is shown in Fig. 9, which shows that the Metric of BRISK features has the most importance in the random forest algorithm.

a) Performance of classification: The performance of all the above classifiers can be measured based on their classification report in terms of training and testing results [3]. The performance can be measured based on four parameters accuracy, precision, recall, and f1-score on the testing results. All these parameters are measured with true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) from the confusion matrix [3]. The formula for every parameter is shown in Table III. This table shows the formula for terms used in the classification report.The histogram is plotted only for f1-score because it is the combination of both precision and recall. The accuracy comparison of the f1-score for all classifiers is shown in Fig. 5.

TABLE III. CLASSIFICATION REPORT PARAMETERS FORMULA

Eq. No.	***Features Type***	***Features Formula***
13	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
14	Precision	$\frac{TP}{TP + FP}$
15	Recall	$\frac{TP}{TP + FN}$
16	F1-Score	$2 * \frac{Precision * Recall}{Precision + Recall}$

E. K-Fold Cross Validation

Cross-validation is a machine learning algorithms technique which mostly used to test the machine learning models are performing effectively. In the case of the limited dataset, the cross-validation can also be used as resampling to evaluate a model [14]. In this case, K-fold cross-validation has been performed on the training dataset taking 10 folds for the confirmation that all the created classifiers have not been overfitted [3]. In K-Fold the process repeats itself for k times so there can be k times Mean Square Error (MSE), and equation (13) shows the formula for MSE. All the accuracy results have been achieved with 10 fold cross-validation; the comparison histogram is shown in Fig. 6.

$$cv_{(k)} = \frac{1}{k} \sum_{\mathrm{i=1}}^{\mathrm{k}} MSE_i \qquad (17)$$

F. Pricncipal Component Analysis (PCA)

PCA is an unsupervised machine learning algorithm used for dimension reduction in the case of a large number of dimensions or features. It shows better accuracy results after reducing the dimension of features of the original dataset because the models with high dimensions or a huge number of features can perform very slowly and most of the time fail to perform classification [6]. PCA is also used to remove the overfitting in classifier models and it also improves the performance of model accuracy at a very low cost [20]. In this case, PCA is also applied because there are total 11 number of features and it is difficult for a model to make the decision so from these 11 only 6 PCA components have been taken for the classification purposes and the accuracy results for the 6 components are comparatively similar to the results before PCA. This proves that reducing the dimension or the number of features gives almost the same accuracy as without PCA. For comparison purposes, the computed accuracy for PCA 6 and 7 components, and all the comparison plots are given in the results section, shown in Fig. 7 and Fig. 8.

G. Transfer Learning

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task [27]. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in a skill that they provide on related problems. Transfer methods tend to be highly dependent on the machine learning algorithms being used to learn the tasks and can often simply be considered extensions of those algorithms [17]. In transfer learning, the initial steps of image acquisition and image preprocessing are the same as shown in Fig. 3, which are applied for traditional classifiers. Data augmentation is a strategy that enables us to significantly increase the diversity of data available for training models without actually collecting new data [23] [24]. In this process, the dataset has been divided into an augmented and unaugmented form which has been further passed for transfer learning techniques, such as cropping, padding, and horizontal flipping are used to augment the data to train a large neural network with small dataset.

H. VGG-16

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman [25], the model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. The model of VGG-16 is shown in Fig. 4, it includes 13 Convolutional layers, 5 pooling layers, and 3 dense/fully connected layers.

a) Convolutional layer: The Convolutional layer is the building block of the neural network; it is application of a filter to an input that results in an activation. A feature map is generated with the repeated application of the same filter in a map of activations, indicating the locations and strength of a detected feature in an input, such as an image [26].

b) Pooling layer: The pooling layer is placed right after the convolutional layer, it provides downsampling of feature maps by summarizing the presence of features in patches of the feature map. Average pooling and max pooling are two common methods that summarize the average presence of a feature and the most activated presence of a feature respectively [22][27].

c) Fully connected layer: Fully connected layers are an essential component of Convolutional Neural Networks (CNNs), which have been proven very successful in recognizing and classifying images for computer vision. The CNN process begins with convolution and pooling, breaking down the image into features, and analyzing them independently. The result of this process feeds into a fully connected neural network structure that drives the final classification decision [27].

d) Softmax/sigmoid layer: The Softmax function is sometimes called multi-class logistic regression because the softmax is a generalization of logistic regression that can be used for multi-class classification, whereas the sigmoid function is used for logistic regression or binary classification [27].

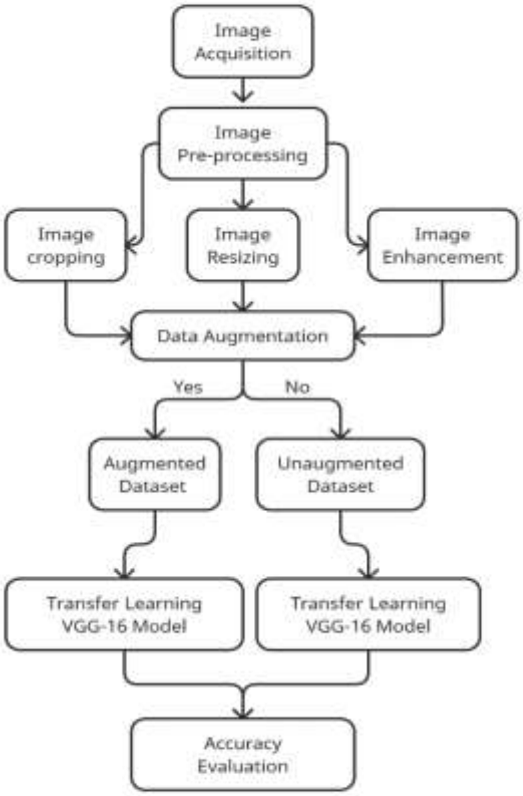


Fig. 3. Proposed Work Flow Chart for Transfer Learning.

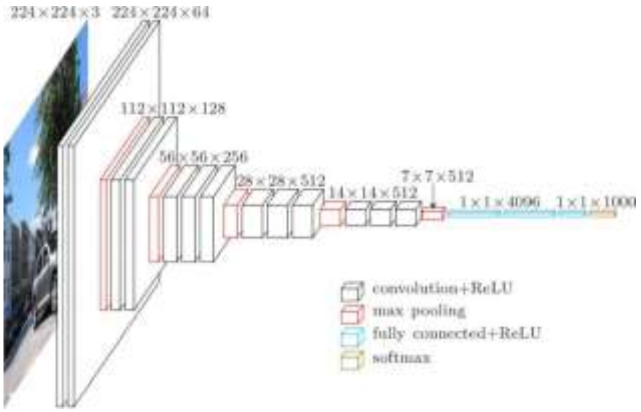


Fig. 4. VGG-16 Model which Includes 13 Convolutional, 5 Pooling and 3 Dense/Fully Connected Layers.

7) Regularization: Dropout is a regularization method that approximates training a large number of neural networks with different architectures in parallel [22]. During training, some number of layer outputs are randomly ignored or "dropped out." This has the effect of making the layer look-like and be treated-like a layer with a different number of nodes and connectivity to the prior layer. In effect, each update to a layer during training is performed with a different "view" of the configured layer [27]. The dropout layer is used in the transfer learning VGG16 model after the relu activation layers to randomly drop the weights and generalize better to remove the overfitting.

IV. Results and Discussion

In this section, all the results of classification accuracy and report have been discussed thoroughly. A comparative analysis has been taken for every classifier's performance and accuracy results before PCA and after PCA. All the resulting histograms are discussed in this section.

A. F1-Score Classification Analysis

The accuracy of all the above discussed classifiers is shown in Fig. 5. The given F1-score comparison plot is the average of both healthy and unhealthy rice crops. F1-score is the combination of precision and recall, so here, F1-score values for the SVM and random forest classifiers with an image size of 256 x 256 and 512 x 512 are higher than other classifiers. It proves that for the fungal blast disease in rice seed, the classification with SVM and the random forest is much better than other classifiers. The random forest has better results for the image size of 256 x 256. F1-score values for the Naïve Bayes classifier are lowest than other classifiers, but for the case of 256 x 256 image size, it has a high score. The remaining classifier has an almost related F1-score, so from this comparison histogram, it has been concluded that based on precision and recall results, the SVM and random forest, both classifiers have higher results and can perform better for the image size of 256 x 256.

B. Classification Accuracy Analysis before PCA

The accuracy comparison histogram before PCA is shown in Fig. 6. Based on the healthy and unhealthy dataset for the fungal blast disease, the final results are shown in the histogram conclude that the accuracy of SVM and random forest classifiers are higher than other classifiers. Six different classification models are applied for 10 folds cross-validation on 80% of the training and 20% of testing of the dataset. The classification has been performed on all image sizes of the dataset and it has been observed that the results for 256 x 256 image size are most accurate. The accuracy results that are achieved with 256 x 256 image size for 10-fold cross validation before PCA are given as under:

The SVM classifier has better performance with testing accuracy of 73.50%. Random Forest classifier also has best performance of 74.80% of testing accuracy. LDA classifier performed well with an accuracy of 70.55%, which is less as compared to SVM and random forest. Logistic regression classifier has an accuracy of 68.05%, which is less than LDA. The Decision Tree classifier has got an accuracy of 65.55%, which is smaller than logistic regression. While, Naïve Bayes achieved an accuracy of 62.33%, which is the lowest as compared to all classifiers, shown in Fig. 6. From the above classification results, Naïve Bayes classifier has very low accuracy while LDA and logistic regression have almost the same accuracies. The decision tree classifier also has good performance, but the results are smaller than SVM and random forest. Finally, it has been concluded that the SVM and random forest both classifiers have better performance with 10 folds cross-validation, and the accuracy is almost 73.50% and 74.80%.

C. Classification Accuracy Analysis after PCA

The PCA classifier is used to reduce the dimension or the features to get better results. In this case, PCA is also applied to reduce the number of features. PCA is applied with 10-fold cross-validation for 6 and 7 components with the reduction of 5 and 4 dimensions from 11 dimensions (features), shown in Fig. 4 and 5. From the comparison histogram before PCA, it can be seen that the results are good, but possibly due to the huge number of features, models get confused and did not perform well. So, here 6 and 7 PCA components are taken, which help the models to for a better decision. The comparison histogram after applying PCA is shown in Fig. 4 and 5. After applying PCA to every classifier similarly, 10 folds cross-validation has been applied for the removal of overfitting.

The accuracy results that are achieved with 256 x 256 image size for 10-fold cross validation for 6 PCA components are given as under: The SVM classifier has better performance with a testing accuracy of 69.03%. Random Forest classifier also has the best performance of 72.52% of testing accuracy. LDA classifier performed well with an accuracy of 68.55%, which is less as compared to SVM and random forest. Logistic regression classifier has an accuracy of 68.05%, which is less than LDA. The Decision Tree classifier has got an accuracy of 67.88%, which is smaller than logistic regression. While, Naïve Bayes achieved an accuracy of 65.53%, which is the lowest as compared to all classifiers, shown in Fig. 6. From the above classification results for 6 PCA components, it has been concluded that the testing accuracy for random forest classifier is higher than all others. So, for 6 PCA components, random forest classifier has better performance.

The accuracy results that are achieved with 256 x 256 image size, for 10-fold cross validation for 7 PCA components are given as under: The SVM classifier has better performance with testing accuracy of 71.45%. Random Forest classifier also has the best performance of 70.65% of testing accuracy. LDA classifier performed well with an accuracy of 68.67%, which is less as compared to SVM and random forest. Logistic regression classifier has a accuracy of 69.08%, which is less than LDA. The Decision Tree classifier has got an accuracy of 67.18%, which is smaller than logistic regression. While, Naïve Bayes achieved an accuracy of 66.12%, which is the lowest as compared to all classifiers, shown in Fig. 6. From the above classification results for 7 PCA components, it has been concluded that the testing accuracy for SVM classifier is higher than all others. So, for 7 PCA components, SVM classifier has better performance.

From overall results, it has been concluded that the accuracy for 6 and 7 components is almost similar to the accuracy for all components. It proves that after reducing the number of features in PCA almost the same results are achieved as before applying PCA. The dimensionality is reduced up to 4 and 5 features. So, both results before PCA and after PCA are almost the same and with the help of these traditional classifiers, maximum achieved accuracy is 75%.

D. VGG-16 Performance Analysis

The VGG-16 classifier is used to classify healthy and unhealthy images, it has been used in two conditions, without data augmentation and regularization, and with data augmentation and regularization. Both VGG-16 classifiers have top layers disabled, and a new model has been created using the pre-trained weights of VGG-16.

The results of VGG-16 without data augmentation and regularization are shown in Fig. 10, the validation accuracy is a maximum of 64%, and it has stopped learning, which indicates that the model is overfitting. Two techniques are used to reduce overfitting in the model, i.e., augmentation and dropout. In data augmentation, the data is increased artificially for the model to learn better in training epochs and regularization, and then dropout regularization is used in which the model randomly drops learned weights after every epoch, which helps the model not to become general. In Fig. 11, it can be seen that the validation accuracy of the model has increased to 71.28% after applying the data augmentation and dropout regularization technique. These techniques play a crucial part in the fine-tuning of the model to achieve the best results.

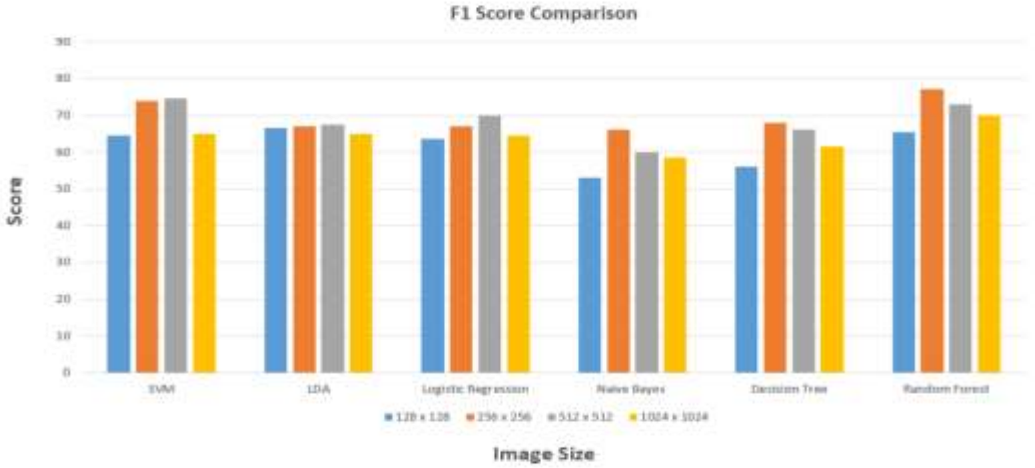


Fig. 5. Comparison Histogram of Every Classifier Discussed above for Classification Report Parameters, F1-Score which is Combinaion of Precision and Recall, this Histogram shws that F1-Score Values for SVM and Random Forest Classifier are Higher.

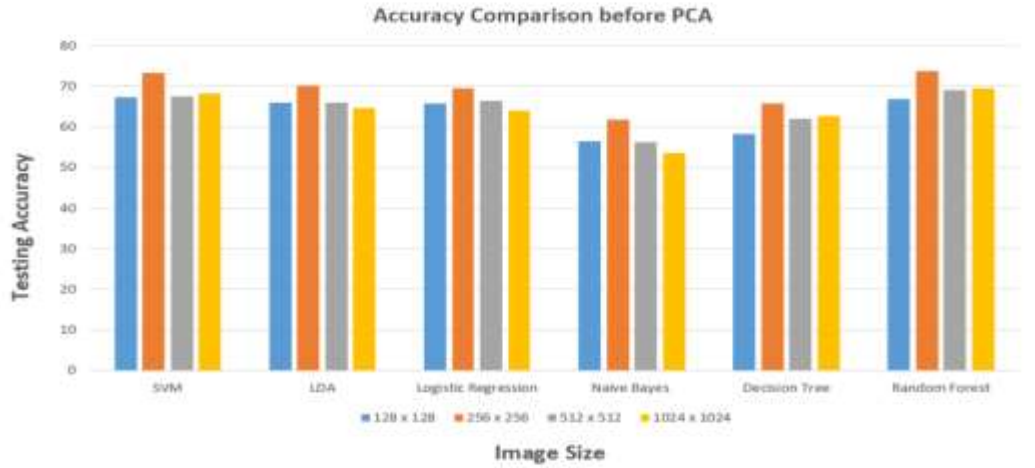


Fig. 6. Comparison Histogram of Every Classifier for Every Size of Image before PCA which Shows that SVM and Random Forest Classifiers have Higher Accuracy for the Image Size of 256 x 256.

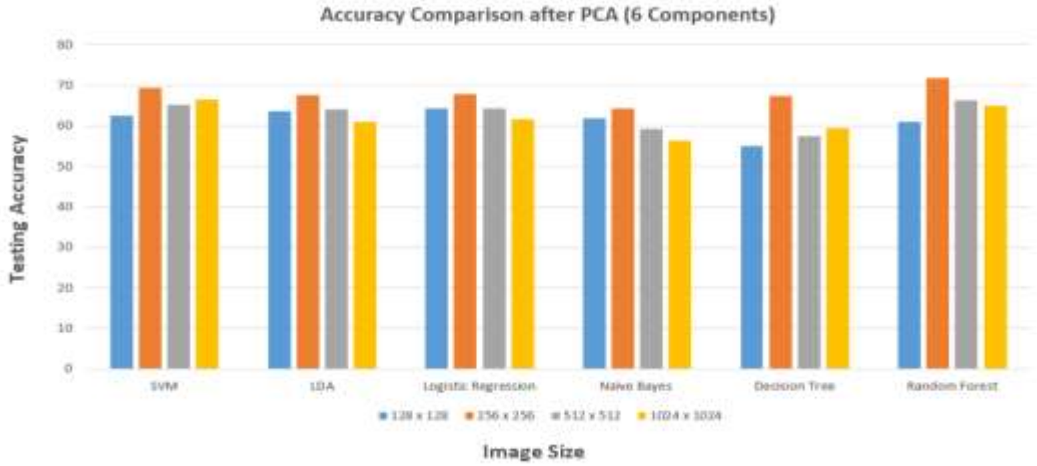


Fig. 7. Comparison Histogram of Every Classifier after PCA for 6 PCA Components which shows that SVM And random Forest Classifiers have Higher Accuracy for the Image Size of 256 x 256.

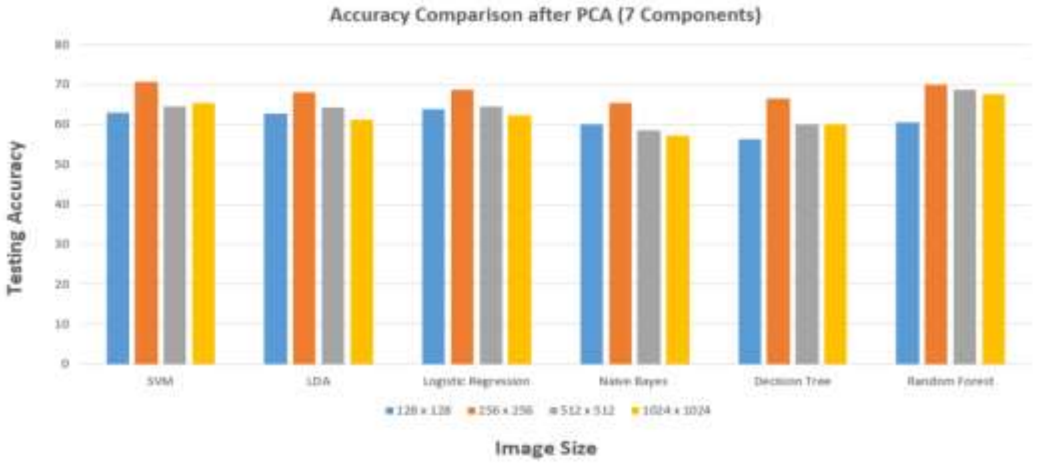


Fig. 8. Comparison Histogram of Every Classifier after PCA for 7 PCA Components which Shows that the Accuracy Results for SVM and Random Forest are Higher for Image Size of 256x256.

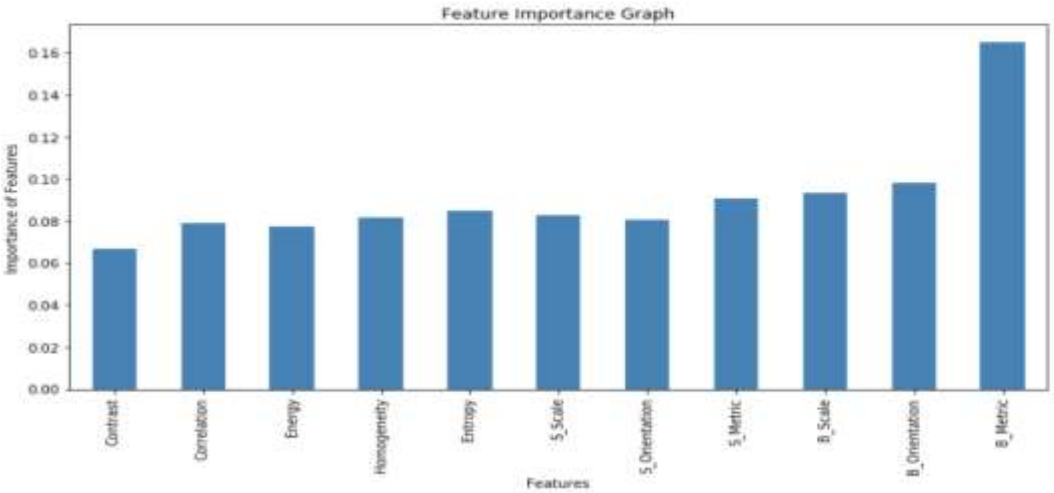


Fig. 9. Feature Importance Histogram for Random Forest Classifier where B and S in Features Axis Stands for BRISK and SURF which Shows that Metric of BRISK Features is the Most Important Feature in Random Forest Classifier.

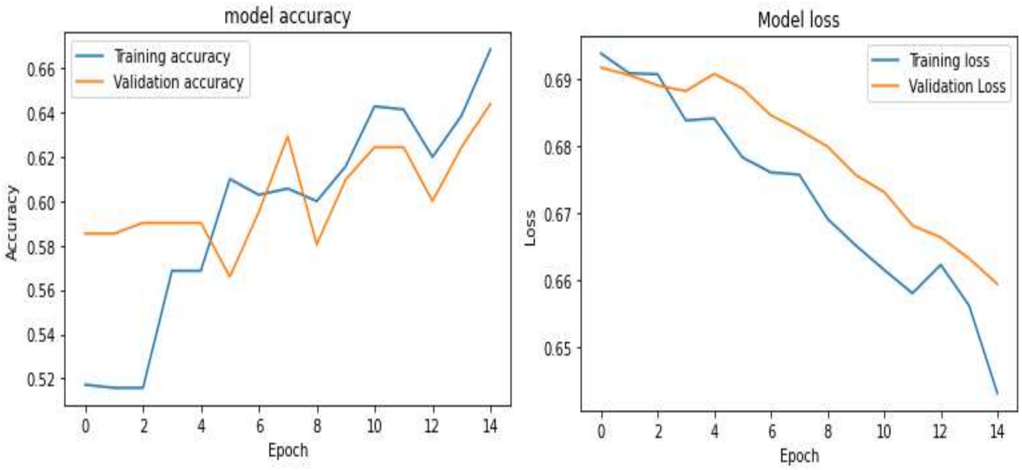


Fig. 10. Training Accuracy and Validation Accuracy along with Training Loss and Validation Loss of VGG-16 Model using Unaugmented Dataset and no Regularization.

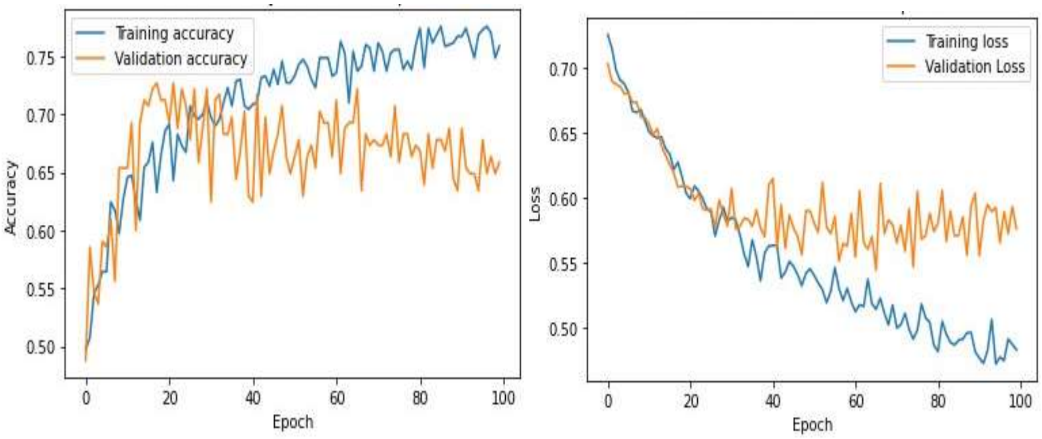


Fig. 11. Training Accuracy and Validation Accuracy along with Training loss and Validation Loss of VGG-16 Model using Augmented Dataset and Regularization.

V. Major Contribution

In this comprehensive research, the major contribution is the unique dataset of rice crops which has been collected from different fields of Kashmore, Pakistan. There are many publications for plant disease detection in general, but regarding the rice plant diseases, limited research work is done that is only for rice leaf diseases. In this research, the fungal blast disease has been detected on rice crop seed with different image processing techniques and machine learning algorithms, which is another major contribution.

VI. Conclusion

Plant disease detection plays an essential role in the growth of the economy and healthy crop production. In the proposed work, the fungal blast disease is detected in the seed of rice crop. This paper discussed the different image processing and machine learning techniques to detect fungal blast disease in rice crops. Image processing is used for the extraction of multiple features and extracted 11 different features from the models such as texture, SURF, and BRISK. As per this research, the mentioned features are beneficial for the detection of fungal blast disease, in which rice has brownish spots on its seed, shown in Fig. 2. In the machine learning portion, a comparative analysis regarding different machine learning algorithms based on disease detection with varying accuracies has been made. Seven different classifiers are used, including traditional and convolutional classifiers. After analyzing these traditional features and classifiers, the dataset has been used as input to transfer learning VGG-16 model, then trained the model with the unaugmented dataset and augmented dataset. After training, the validation accuracy of the trained model with the unaugmented dataset was 64%, while the accuracy of the trained VGG-16 model with the augmented dataset was 71.28%.

Finally, it has been concluded that after applying PCA with 10-fold cross validation, the random forest algorithm has still the best performance for the fungal blast disease detection with an accuracy of 73.12% for the testing dataset in the traditional classifiers whilst the highest accuracy from transfer learning dataset is of 71.28%. If analyzed, it is not a big difference as compared to the efforts that have been put in order to run the traditional classifiers while the images data was input to the transfer learning model.

VII. Future Scope

In future work, the plan is to develop a mobile application and an agricultural cultivating drone for fungal blast disease detection in rice crop seed during the field. This mobile application will help farmers to detect the disease in rice seed by capturing an image of the plant in the field, and they will get the most accurate and fast results on the spot. Similarly, an agricultural drone will visit the whole field and will monitor the plant's health. Based on those results of drone and mobile applications, the farmers can use related pesticides and fertilizers to improve the health of the crop. This technology will reduce the cost for extra use of pesticides, and farmers will get a good profit while giving only the needed pesticides to crops, it will be more beneficial for the economy of this country.

References

[1] S. D. Khirade and A. B. Patil, "Plant Disease Detection Using Image Processing," 2015 International Conference on Computing Communication Control and Automation, Pune, 2015, pp. 768-771, doi: 10.1109/ICCUBEA.2015.153.

[2] P. Panchal, V. C. Raman and S. Mantri, "Plant Diseases Detection and Classification using Machine Learning Models," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-6, doi: 10.1109/CSITSS47250.2019.9031029.

[3] U. B. Korkut, Ö. B. Göktürk and O. Yildiz, "Detection of plant diseases by machine learning," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404692.

[4] Raut Sandesh, Fulsunge Amit, Plant Disease Detection in Image Processing Using MATLAB, 2017, Volume-6, ISSN 2319-8753.

[5] O. Kulkarni, "Crop Disease Detection Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697390.

[6] L. S. Puspha Annabel, T. Annapoorani and P. Deepalakshmi, "Machine Learning for Plant Leaf Disease Detection and Classification – A Review," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0538-0542, doi: 10.1109/ICCSP.2019.8698004.

[7] K. Ahmed, T. R. Shahidi, S. M. Irfanul Alam and S. Momen, "Rice Leaf Disease Detection Using Machine Learning Techniques," 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/STI47673.2019.9068096.

[8] "Rice leaf diseases data set." https://archive.ics.uci.edu/ml/datasets/Rice+Leaf+Diseases. Accessed: 2019-09-27.

[9] Kurale, Neha & Vaidya, Madhav. (2018). Classification of Leaf Disease Using Texture Feature and Neural Network Classifier. 1-6. 10.1109/ICIRCA.2018.8597434.

[10] Anjna, Sood, M., & Singh, P. K. (2020). Hybrid System for Detection and Classification of Plant Disease Using Qualitative Texture Features Analysis. Procedia Computer Science, 167(2019), 1056–1065. https://doi.org/10.1016/j.procs.2020.03.404

[11] S. Ramesh *et al*., "Plant Disease Detection Using Machine Learning," *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, Bangalore, 2018, pp. 41-45, doi: 10.1109/ICDI3C.2018.00017.

[12] P. Sharma, P. Hans and S. C. Gupta, "Classification Of Plant Leaf Diseases Using Machine Learning And Image Preprocessing Techniques," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 480-484, doi: 10.1109/Confluence47617.2020.9057889.

[13] R. Naga Swetha, V. Shravani, Monitoring of Rice Plant for Disease Detection using Machine Learning, Volume-9, ISSN 2249-8958, https://www.ijeat.org//papers/v9i3/C5308029320.pdf.

[14] Muhammad Kashif, Thomas M. Deserno, Daniel Haak, Stephan Jonas, Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? A general question answered for bone age assessment, Computers in Biology and Medicine, Volume 68, 2016, Pages 67-75, ISSN0010-4825, https://doi.org/10.1016/j.compbiomed.2015.11.006.

[15] M. A. Jasim and J. M. AL-Tuwaijari, "Plant Leaf Diseases Detection and Classification Using Image Processing and Deep Learning Techniques," 2020 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, 2020, pp. 259-265, doi: 10.1109/CSASE48920.2020.9142097.

[16] Prajapati, Harshadkumar & Shah, Jitesh & Dabhi, Vipul. (2017). Detection and classification of rice plant diseases. Intelligent Decision Technologies. 11. 357–373. 10.3233/IDT-170301.

[17] U. Shruthi, V. Nagaveni and B. K. Raghavendra, "A Review on Machine Learning Classification Techniques for Plant Disease Detection," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 281-284, doi: 10.1109/ICACCS.2019.8728415.

[18] E. Hossain, M. F. Hossain and M. A. Rahaman, "A Color and Texture Based Approach for the Detection and Classification of Plant Leaf Disease Using KNN Classifier," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ECACE.2019.8679247.

[19] B. S. Kusumo, A. Heryana, O. Mahendra and H. F. Pardede, "Machine Learning-based for Automatic Detection of Corn-Plant Diseases Using Image Processing," 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, Indonesia, 2018, pp. 93-97, doi: 10.1109/IC3INA.2018.8629507.

[20] L. S. Puspha Annabel, T. Annapoorani and P. Deepalakshmi, "Machine Learning for Plant Leaf Disease Detection and Classification – A Review," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0538-0542, doi: 10.1109/ICCSP.2019.8698004.

[21] Kumar, D. S. and Samrity. "Plant Species Identification using SIFT and SURF Technique." (2017), ISSN 2319-7064, https://www.ijsr.net/archive/v6i3/ART20171974.pdf.

[22] Arsa, D. M. S., & Susila, A. A. N. H. (2019). VGG16 in Batik Classification based on Random Forest. *Proceedings of 2019 International Conference on Information Management and Technology, ICIMTech 2019*, *1*(August), 295–299. https://doi.org/10.1109/ICIMTech.2019.8843844.

[23] Weiss, K., Khoshgoftaar, T. M., & Wang, D. D. (2016). A survey of transfer learning. In *Journal of Big Data* (Vol. 3, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-016-0043-6

[24] Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *ArXiv*.

[25] Khan, Ufaq. "Plant Disease Detection Techniques: A Review." (2019), IJCSMC (Vol. 8, No. 4), Pages 59-68, http://paper.researchbib.com/view/paper/206987.

[26] N.K Ambika, P Supriya, Detection of Vanilla Species by Employing Image Processing Approach, Procedia Computer Science, Volume 143, 2018, Pages 474-480, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.420.

[27] Sharma, P., Hans, P., & Gupta, S. C. (2020). Classification of plant leaf diseases using machine learning and image preprocessing techniques. Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering, 480–484. https://doi.org/10.1109/Confluence47617.2020.9057889