Development of an intelligent laser biospeckle system for early detection and classification of soybean seeds infected with seed-borne fungal pathogen (Colletotrichum truncatum)

Article *in* Biosystems Engineering · November 2021

DOI: 10.1016/j.biosystemseng.2021.11.002

CITATIONS
0

READS
25

8 authors, including:

Puneet Singh Thakur
Indian Institute of Technology Indore
28 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Amit Chatterjee
Independent Researcher
53 PUBLICATIONS **117** CITATIONS
SEE PROFILE

Laxman singh Rajput
Directorate of Soybean Research
37 PUBLICATIONS **89** CITATIONS
SEE PROFILE

BIOSYSTEMS ENGINEERING 212 (2021) 442–457

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Development of an intelligent laser biospeckle system for early detection and classification of soybean seeds infected with seed-borne fungal pathogen (*Colletotrichum truncatum*)

Puneet Singh [a], *Amit Chatterjee* [a], *Laxman S. Rajput* [b], *Santosh Rana* [c], *Sanjeev Kumar* [b], *Vennampally Nataraj* [b], *Vimal Bhatia* [a], *Shashi Prakash* [c,*]

[a] *Signals Discipline of Electrical Engineering & Centre for Advance Electronics, Indian Institute of Technology, Indore, 453552, India*

[b] *ICAR- Indian Institute of Soybean Research, Indore, 452001 India*

[c] *Photonics Laboratory, Department of Electronics & Instrumentation Engineering, Institute of Engineering & Technology, Devi Ahilya University, Khandwa Road, Indore, 452001, India*

ARTICLE INFO

Article history:
Received 2 January 2021
Received in revised form
18 October 2021
Accepted 1 November 2021
Published online 18 November 2021

Keywords:
Agriculture
Biospeckle Analysis
Early Detection
Laser
Machine Learning
Seed-Borne Fungi

There is a need for developing rapid and non-destructive techniques for the early detection of seed-borne fungal pathogen because they can be an essential step towards adopting effective disease control measures. Existing techniques for detecting seed-borne diseases have poor sensitivity towards early stages of pathogen development (i.e., when seeds are asymptomatic) and they are also expensive, time-consuming, complex, require mycological skills and destructive testing operations. Aiming at overcoming the above limitations of the existing techniques, a novel laser biospeckle based method is proposed for early detection of seed-borne fungal infection in conjunction with machine learning. Soybean seeds infected by low concentrations (10^2-10^6 spores ml^{-1}) of *Colletotrichum truncatum* were analysed by using full field biospeckle analysis to establish the possible relationship between biological activity in early stages of pathogen infection, with and without the use of frequency filtering. The results demonstrate that the biospeckle activity (BA), for both, raw and frequency filtered data was significantly high ($p < 0.05$) for the diseased seeds even for low inoculum concentrations. Moreover, the amplitude values of mid frequency spectral components for diseased seeds were higher than those of lower and higher spectral components which correspond to the BA of fungal infected seeds. Several classical machine learning algorithms were trained to model the response of healthy and diseased samples after parameter optimisation. Obtained results showed that k-nearest neighbour (k-NN), decision tree (DT), and artificial neural network (ANN) based predictive models presented strong robustness and high performance with overall accuracy reaching up to 96.94% for classifying diseased seeds.

* *Corresponding author*. Fax: +91 731 2764385.
E-mail address: sprakash_davv@rediffmail.com (S. Prakash).

https://doi.org/10.1016/j.biosystemseng.2021.11.002
1537-5110/

Nomenclature

Abbreviations

ACD	Area covered by disease
AD	Average difference
ANNs	Artificial neural networks
ANOVA	Analysis of variance
AU-ROC	Area under the ROC curve
BA	Biospeckle activity
BOD	Bio-oxygen demand
CCD	Charge-coupled device
COM	Co-occurrence matrix
CRD	Completely randomised design
DNA	Deoxyribonucleic acid
DT	Decision tree
FPR	False positive rate
HIS	Hyperspectral imaging
HSD	Honest significant difference
HSI	Hue-saturation-intensity
IM	Inertia moment
k-NN	k-Nearest neighbour
LR	Logistic regression
MO	Microscopic objective
NB	Naive Bayes
NIRS	Near-infrared spectroscopy
PCR	Polymerase chain reaction
QTP	Quality test protocols
RBF	Radial basis function
ReLU	Rectified linear unit
ROC	Receiver operating characteristics
ROI	Region of interest
SVM	Support vector machine
THSP	Time history of speckle patterns
TP, FP, TN, FN	True positive, false positive, true negative, false negative
TPR	True positive rate

Symbols (units)

C	Regularisation
I	Sequence of speckle images
k	Number of neighbours in k-NN
N	Number of time sequence speckle images
Np, Nq	Dimensions of speckle images
n	Number of features
n_s	Number of average spores (ml^{-1})
p	p-value of the pairwise comparison
x_i, y_i	Elements of feature vector

1. Introduction

Nearly all the existing food crops in the world are produced through seeds. Growing healthy and clean seeds is the first and foremost method of plant disease management. Nevertheless, the production of healthy and clean seeds is a serious challenge due to the existence of seed-borne pathogens. Apart from reducing the quantity and quality of harvested seed, seed-borne pathogens remain preserved in seed lots, which can enormously increase the spread of plant pathogens (Mancini & Romanazzi, 2014). In most cases, seeds do not show any visible symptoms of being infected unlike other plant tissues (Rajput et al., 2020; Schaad et al., 2003) and thus infection go unhindered.

Over the last few decades, several methods have emerged to ascertain the existence of seed-borne pathogens. However, visual examination, selective growth media, serological assays, and bioassay have commonly been used for pathogen detection (Kumar & Gupta, 2020). But these methods are inefficient, time consuming and possess less specificity and sensitivity towards seed-borne pathogen detection. Other molecular diagnostics methods (viz. polymerase chain reaction (PCR), multiplex PCR, Bio-PCR, DNA barcoding etc.) possess vast potential for improving pathogen detection in seeds. However, these methods are expensive, time-consuming and destructive to the test samples, making them impossible for large scale non-destructive screening or integration in an on-line sorting and production system. Optical methods such as fluorescence spectroscopy, near-infrared spectroscopy (NIRS) and hyperspectral imaging (HSI) have also been used for non-destructive detection of aflatoxins (mycotoxins) and fungal infection in wide varieties of agricultural products (Tao et al., 2018). However, the above discussed methods are not sensitive to the low concentrations of fungi spores and they do not provide accurate result when the seeds are asymptomatic. Nearly 100 spores seed^{-1} can establish a typical disease on the seeds of any kind of susceptible variety (Rajput et al., 2020), but these methods required more than 1000 spores seed^{-1} for detection of a pathogen. Hence, early identification of infection with low spore concentration in seeds is not possible by using these methods.

Since methods associated with the detection of seed-borne pathogens are the first line approach for the supervision and mitigation of diseases, rapid detection of the pathogen at early stages of its progression becomes essential to satisfy the requirements of effective disease control. In the current study, an attempt is made for automatic identification of fungal pathogen at the early stages of its development (i.e. when seed does not show any symptom of the disease) by using a laser biospeckle technique. With laser biospeckle, a coherent light source irradiates a sample having some physical or biological activity (Zdunek et al., 2014). Biospeckle activity (BA) of the sample is evaluated by recording the time sequence of light intensity across the illuminated samples, and performing image processing to retrieve useful information. The technique has been extensively used as a non-destructive measurement tool, to monitor biospeckle/biological activity of the samples in agriculture (Braga et al., 2005, Rabelo et al., 2011, Singh et al., 2020b; Zdunek et al., 2014), engineering (Kooij et al., 2016), pomology (Singh et al., 2018; Zdunek & Cybulska, 2011), biomedical imaging (Carvalho et al., 2009; Chatterjee et al., 2018a, 2018b; Singh et al., 2020a), and biometrics (Chatterjee et al., 2017, 2018a, 2018b, 2019).

Only a few studies have investigated the potential of laser biospeckle technique for screening the fungi species in bean seeds (Braga et al., 2005; Rabelo et al., 2011). Braga et al. (2005) developed biospeckle based method for detecting the

presence of fungal infection in bean seeds by using visual method for qualitative assessment as well as single column based numerical processing for quantitative measurements. Extending the notion of the given work, Rabelo et al. (2011) presented a biospeckle technique for identification of fungi species that was complemented by a frequency domain analysis of the acquired data. However, to the best of authors knowledge, the use of laser biospeckle technique for early identification of fungus pathogen in seeds by using full field frequency domain analysis and machine learning based automatic classification of infected seeds has not yet been attempted.

Moreover, in the aforementioned works (Braga et al., 2005; Rabelo et al., 2011), the authors have endeavoured to find the difference in the quantitative values of biospeckle activity between the healthy and infected seeds by using single column based approach (co-occurrence matrix (COM) in conjunction with inertia moment (IM) technique) for numerical quantification. However, these quantification strategies have several critical limitations. These single columns based techniques utilised manual region of interest (ROI) selection, which increases the memory requirement and may generate ambiguous results (as the pixel values of the middle column may alter due to manual selection) (Chatterjee et al., 2020). Moreover, previous analyses were performed by considering the fungal growth on biospecimen to be homogeneous; hence, time history of speckle patterns (THSP) was generated by using only a single column from each speckle frame. However, in real life, these phenomena are inhomogeneous (Cardoso et al., 2011). It is a well-established fact that the seeds are heterogeneous in nature and interaction of seeds with fungal microorganism made it somehow more complex, and hence measurements may produce large standard deviations (due to considering a single column) and this decreases the reliability of indices.

In this work, to circumvent all the limitations of the above discussed processing strategies, average difference (AD) is utilised (Dai Pra et al., 2016). It is based on a full field visual assessment and numerical indexing for assessment of fungal activity in the seed specimen. AD based full field indexing techniques utilise every pixel of each speckle frame for biospeckle activity assessment and they are advantageous over previously used numerical technique due to its zero standard deviation, higher accuracy, and efficient handling of both homogeneous as well as heterogeneous activities (Chatterjee et al., 2020). Moreover, BA is the sum of various physiological and biochemical processes like biochemical reactions, cytoplasmic streaming, cell divisions, organelle movement, Brownian motion, and many more (Zdunek et al., 2014) occurring inside a biological sample. However, the traditional methods of analysis lack the ability to separate or isolate certain biospeckle features associated with a specific biological phenomenon. Hence, frequency decomposition is also carried out (Alves et al., 2013; Braga et al., 2007; Cardoso et al., 2011; Nobre et al., 2009, Sutton and Punja, 2017) on biospeckle data to extract activity associated with seed-borne fungus pathogen for early identification of disease symptoms.

Machine learning approaches have brought new and promising perspectives in the field of agriculture and have proven their efficiency in several applications. Automatic detection and recognition of fungus pathogen in symptomless seeds by using biospeckle technique in conjunction with machine learning is useful in identifying the disease at early stage of its development. One of the objectives of this research is focused on analysing the performance of different variants of supervised machine learning algorithms in early identification of disease symptoms in seeds. Prediction of diseases in plants and seeds, have recently gained significant attention due to the wide adaptation of computer based technology into the agricultural sector. Researchers have achieved high accuracy when classifying various plant diseases induced by fungal, viral and bacterial pathogens using machine learning based classifiers (Rumpf et al., 2010). The combination of these machine learning algorithms with optical techniques can overcome several limitations faced by the traditional methods of seed inspection.

In this context, demand for developing a rapid, non-destructive and automatic technique for real-time detection of contamination in seeds has received significant attention. Among currently emerging technologies the optical-based techniques have been reported to show great potential for real-time applications in this direction (Tao et al., 2018). Therefore, this investigation outlines the foundational research to facilitate the issue of automatic identification of seed-borne fungal disease in symptomless soybean seeds during early stages of pathogen development. The main objectives of the study were (i) to develop a method based on biospeckle analysis to detect seed-borne fungus pathogen during early stage of infection development in asymptomatic seeds, (ii) to analyse the progression of pathogen infection in seeds using AD based full field analysis technique by considering the optical inhomogeneity present in the sample (iii) to perform frequency decomposition of the biospeckle data for isolating the biological phenomena associated with the pathogen infection, and (iv) to develop a machine learning based automatic classification approach for real-time identification and classification of diseased seeds.

2. Materials and methods

2.1. *Preparation of biological samples*

2.1.1. *Culture of pathogenic agent* (Colletotrichum truncatum)

Several seed-borne fungi have been reported to produce anthracnose disease in soybean. However, anthracnose caused by *C. truncatum* is considered as the most predominant disease in Central India (Nataraj et al., 2020). In this study, *C. truncatum* was isolated from typical diseased pod samples collected from plant protection field of ICAR-Indian Institute of Soybean Research (IISR), Indore (India). Collected fungi was multiplied on potato dextrose agar culture medium (HiMedia) and incubated at 27 °C (Rajput et al., 2016). After 10 d of incubation *C. truncatum* culture was examined and identified by preparing temporary glass slide under light microscope (Leica). The spore's morphology and size were the main morphological criteria used for identification of *C. truncatum* (Nataraj et al., 2020).

2.1.2. Preparation of spore solutions with different concentrations

Harvesting of fungi was performed by pouring 6 ml of sterile double distilled water into 10 d old sporulated culture petri plate. The petri plate was gently scrapped by using glass rod up to the entire fluffy mycelia. Mycelia was collected into sterile double distilled water and then filtered through muslin cloth. The filtered mycelia were used as a stock for preparing solutions with different spore concentrations of *C. truncatum*. Next, the haemocytometer (Superior Marienfeld, Germany) and coverslips were prepared to count the number of spores for developing solutions with different spore concentrations. Both haemocytometer and coverslips were first sterilized in autoclave at 121 °C for 20 min and then surface sterilized with 70% alcohol for 1 min. Next, 10 μl of spore solution from stock was mixed with 10 μl of 0.4% trypan blue solution and kept exactly over the central depression of the counting chamber of haemocytometer on both the sides. Then, haemocytometer was covered with cover-slips and placed stationary for 5 min for settling down spores. Live spores were counted under the light microscope (Leica) with 100 × magnification by using standard protocol in five replications (Schütz et al., 2020). Thus, spore concentration was calculated by using:

$$\text{Number of spores}\left(ml^{-1}\right) = (n_s) \times 10^4 \quad (1)$$

where, n_s is the average spores in the four 1 mm corner squares of the hemacytometer.

Finally, the stock was diluted serially, and spore concentration was adjusted to 10^2, 10^3, 10^4, 10^5 and 10^6 spores per ml of water by using haemocytometer.

2.1.3. Inoculation of spore on soybean seeds

1500 healthy seeds of JS 20–29 variety were selected from the lot and disinfected by using sodium hypochlorite and distilled water (1:1) for 30 s. Seeds were then rinsed with autoclaved distilled water and dried at room temperature on sterilised sheets of filter paper. The seeds were divided into two groups namely healthy and diseased for experiments. The experiment was conducted as a completely randomised design (CRD) and replicated five times with twenty seeds per replicate. To create an infection, 20 μl of spore solution (with different concentration) was placed uniformly on seeds of each diseased group. Table 1 describes the treatments performed for conducting the experiments. A spore concentration of 10^2 spores ml^{-1} [water] (equivalent to 2 spores $seed^{-1}$) was selected as the lowest concentration to produce symptomless seed for early identification. Furthermore, varying concentrations of inoculum with 20, 200, 2000 and 20000 spores $seed^{-1}$ (10^3, 10^4, 10^5 and 10^6 spores ml^{-1} [water], respectively) were also used to study feasibility of the laser biospeckle technique and the standard protocol for early detection of the pathogen.

For comparison, the healthy group of seeds was treated with 20 μl of sterile distilled water. After treatments of seed with potential inoculums all the samples were kept in biooxygen demand (BOD) chamber at 27 °C for further experimentation.

2.2. Experimental setup and data acquisition

After completion of the inoculation treatments, seeds (healthy and diseased) were removed from the BOD chamber and subjected to dynamicity assessment. A schematic diagram of the experimental arrangement used for recording of the biospeckle data is shown in Fig. 1. The experimental apparatus consists of a He–Ne laser source (15 mW, λ = 632.8 nm), variable attenuator, spatial filtering arrangement, charge-coupled device (CCD) camera, and a personal computer with an image processor. The samples were placed on a vibration isolation tabletop and imaged using the biospeckle apparatus. The intensity of the laser beam was controlled by using a variable absorptive neutral density filter (Thorlabs, USA). These filters reduce the optical power of an incident beam by absorption in the visible and the near IR region. Spatial filtering arrangement consisting of microscopic objective (MO) of magnification 40× and a pin hole of diameter 10 μm, was used to filter and expand the laser beam so that it covers the entire seed sample. Distance between laser source and pinhole of spatial filtering arrangement was 80 mm. Filtering arrangement also reduced non-uniformity generated due to noise in the laser profile to a considerable extent and produced a uniformly illuminated laser beam. Filtered and expanded laser beam is collimated by using a precision achromatic doublet lens (focal length = 250 mm and diameter = 50 mm) and allowed to fall on the mirror placed at an angle of ≈60°, having distance of 330 mm from filtering arrangement. The illumination beam was made to fall on the sample to irradiate the seed. A digital colour CCD camera (Basler Corp., frame rate: 32 frames s^{-1}, resolution: 1024 × 967, Germany) was used to capture the generated biospeckle patterns. Finally, speckle images for healthy and diseased seeds were recorded by using CCD camera positioned perpendicular (approximately 150 mm) to the seed sample with a single frame exposure time 0.03125 s. High quality speckle images were obtained by adjusting iris of the optical lens and magnification of sample space (associated with the relative distance between the sample and the imaging device). Once the optimum quality parameters (Moreira et al., 2014) are set, 128 frames for each stack of biospeckle images (for healthy and diseased groups) were recorded, five times during 48 h, with 12 h intervals between every consecutive recording. After

Table 1 – Description of treatment applied to soybean seeds.

Class	Treatment notation	Inoculums of *C. truncatum* with different spore concentration (spores ml^{-1})	Inoculums of *C. truncatum* with different spore concentration (spores $seed^{-1}$)
Healthy	H	0	0
Diseased	D_2	1×10^2	2
	D_3	1×10^3	20
	D_4	1×10^4	200
	D_5	1×10^5	2000
	D_6	1×10^6	20000

each instance of recording, the seeds were replaced into the BOD chamber at 27 °C.

To validate and compare the results obtained from biospeckle method corresponding to anthracnose disease on soybean seeds, percentage area covered by disease (%ACD) was measured by using ASSESS 2.0 (American Phytopathological Society) software (Madhusudhan et al., 2019; Rajput et al., 2017). ASSESS 2.0 is an interactive laboratory tool for real-time measurement and quantification of disease in seeds and plants by measuring area and length of infected region. The software utilises hue-saturation-intensity (HSI) colour model and perform mathematical transformations to outline diseased area for quantification of disease (Bock et al., 2009). The HSI model is used to separate the seed from the background, and to subsequently separate the diseased area from the seed. The images of healthy and diseased seeds were taken using a uniform background to remove the possibility of manual ROI selection as ASSESS software can easily identify the samples having uniform background. The software provides feature for automatic calculation of the percentage area covered by the disease for quantifying the infection. Figure 2 (a) shows the sample images for healthy seed and Fig. 2 (b) shows the coverage of soybean seed by the disease.

3. Methodology

Figure 3 represents the flow diagram for overall processing pipeline used to process the raw biospeckle data. The processing pipeline consists of three distinct stages of operations: (i) data pre-processing and feature extraction, (ii) predictive model development, and (iii) model evaluation.

3.1. Data pre-processing and feature extraction

The first stage of the pipeline involves pre-processing of the raw biospeckle data and extraction of the associated features for training and testing machine learning models. Firstly, to consider only active data points from the acquired raw biospeckle images region of interest (ROI) selection was performed. A single RGB image of the seed was acquired to generate an image mask for automatic selection of ROI (Chatterjee et al., 2020) from the speckle images of both the groups (healthy and diseased). Figures 4 (a) and (b) show the speckle image captured for soybean seed and its corresponding mask multiplied image, respectively. To acquire good quality speckle images and to reduce subjectivity of results into experimental parameters, the quality test protocols (QTP) were followed (Moreira et al., 2014). Optimum values of quality parameters were obtained and fixed for further experiments. Since, the bioactivity of the specimen changes with time which results in the intensity fluctuations of generated speckle patterns. In order to analyse the physiological or biochemical activity of the samples, N time sequence of intensity images $I = [I_0, I_1, \ldots, I_N]$ of size $N_p \times N_q$ were captured for the time period T with the time interval Δt between consecutive speckle frames. The captured speckle frames were processed by using AD (Dai Pra et al., 2016) based algorithm, with and without frequency filtering. Frequency filtering of raw biospeckle images was performed to isolate specific spectral information associated with several physical or chemical phenomena (e.g. seed water interaction, fungi growth, senescence in seed tissues) (Alves et al., 2013). Finally, BA was determined from raw biospeckle images, and after frequency decomposition the data was used as feature vector for the classification of healthy and diseased seeds.

3.1.1. Frequency decomposition of biospeckle data

For frequency decomposition of biospeckle data, Butterworth function based filter banks were utilised (Sendra et al., 2005). This strategy was preferred over other spectral decomposition methods due to its inherent advantages, like simple design, low complexity, and flat frequency response (Sendra et al., 2005). For frequency filtering, ten 5° order Butterworth filter banks were designed to analyse the time series speckle data. The evolution of intensity in time sequence of consecutive

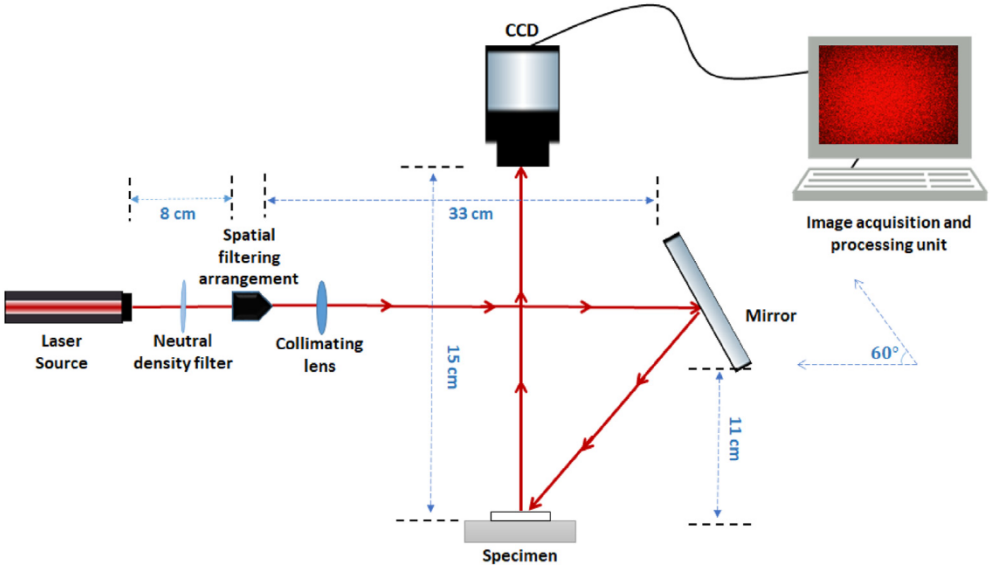


Fig. 1 – Experimental configuration for the backscattering biospeckle data acquisition.

speckle frames is considered as signal and provided as the input of the filter banks. Speckle images were captured using a CCD camera with the time interval of 0.03125 s (i.e. frame fate of 32 frames s^{-1}). The maximum frequency that could be observed, based on the sampling theory, was $1/(2 \times 0.03125)$s which was 16 Hz (Cardoso et al., 2011). To evaluate frequency dependent behaviour of biospeckle data, images were divided into 10 frequency bands by using the generated bank of filters and the width of each frequency band was 1.6 Hz.

3.1.2. *Average difference method*

After decomposition of activity images into different frequency bands, AD (Dai Pra et al., 2016) index corresponding to each sub-band as well as original image stack was evaluated. The calculation of AD activity map is based on the summation of an absolute difference, weighted by local average between pixels of two consecutive speckle frames I_m and I_{m-1}, and is mathematically expressed as:

$$AD(a,b) = \sum_{m=1}^{N} \left| \frac{I_m(p,q) - I_{m-1}(p,q)}{I_m(p,q) + I_{m-1}(p,q)} \right|, \quad (2)$$

where, m is the index for image sequence, N is the number of time sequence speckle images and p and q are the pixel coordinates of image matrix I_m, and $AD(a,b)$ is the resulting 2-D visual activity map.

The size of soybean seeds is not identical; hence, the pixel area of the captured speckle frames was different for each specimen. Hence, to make the resultant biospeckle index invariant to the size of the seeds, spatial averaging of entire activity map was performed (Chatterjee et al., 2020). This is mathematically given by:

$$AD_{index} = \frac{1}{(N_p \times N_q)} \sum_{p=1}^{N_p} \sum_{q=1}^{N_q} \left(\sum_{m=1}^{N} \left| \frac{I_m(p,q) - I_{m-1}(p,q)}{I_m(p,q) + I_{m-1}(p,q)} \right| \right), \quad (3)$$

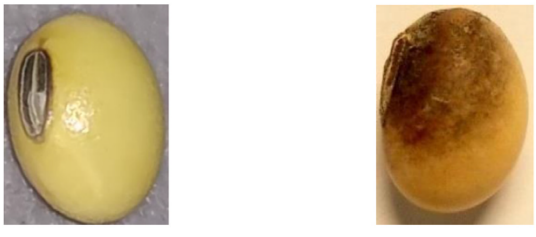
Moreover, variation in the activity value generated due to number of speckle frames used for analysis was nullified by performing temporal frame averaging (Chatterjee et al., 2020). In this step, to make the index independent of number of frames, the generated activity map was divided by total number of frames used for the analysis.

Finally, normalisation of the obtained biospeckle activity was performed by using "Min-Max" normalisation technique (Benhar et al., 2020). In this technique, data normalisation is performed by using linear transformation on the original data. This is the popular normalisation technique as it preserves the relationships among the original data values and removes the bias while comparing measurements having different scales.

Letting X be a numeric data with n observed values, x_1, x_2, …, x_n. "Min-Max" normalisation maps the data value $x_i (i = 1, 2, \ldots, n)$ from population X to a new value x_i' by utilising the maximum and minimum values in the data set. To perform normalisation, minimum and maximum values are recovered from the data and each value is replaced according to the following (Benhar et al., 2020):

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (4)$$

where x_i' is the new value of each entry in data, x_i is the old value of each entry in data X, x_{min} is the minimum and x_{max} is the maximum value of the data to be normalised.



(a) (b)

Fig. 2 – Sample images of soybean seed for (a) healthy (b) diseased groups.

3.2. *Predictive model development*

As shown in Fig. 3, the obtained AD_{index} (for both, frequency filtered and raw data) was used as a feature vector for performing classification task. The dataset resulting from the aforementioned stage of feature extraction (Section 3.1) was partitioned into training, testing and validation datasets for generating learned models. In the training phase, training data set was used for the development of learned models for prediction of the intended class, and during validation phase, the validation dataset was used for unbiased evaluation of predictive models whilst tuning the hyperparameters associated with each learning algorithm. Finally, during the testing phase, test dataset was used for the evaluation of a final model generated from the training dataset. Various state of the art machine learning classifiers, including support vector machine (SVM) (Cortes & Vapnik, 1995), logistic regression (LR) (Larose, 2006), k-nearest neighbour (k-NN) (Bramer, 2007), decision tree (DT) (Polat & Güneş, 2007), Naive Bayes (NB) (Mukherjee & Sharma, 2012), and artificial neural networks (ANNs) (Haykin, 1994) were developed and compared in terms of their feasibility for classifying diseased seeds after optimising their hyperparameters. The data set containing two classes (healthy and diseased seeds), with the level "0" for healthy seeds, and "1" for diseased seeds while considering the training data set consist of n-datum, given as:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \quad (5)$$

where $x \in R^n$ (an n-dimensional space), and $y \in \{0, 1\}$, correspond to two classes.

3.2.1. *Hyperparameters optimisation*

Optimisation of hyperparameters of each learning algorithm have a great influence on performance of the classifier, hence it becomes a crucial step to set optimum values of each hyperparameter to obtain robust and accurate classifiers. The parameters associated with each classification algorithm are summarized in Table 2. Grid-search (Chen & Li, 2010) based optimisation method was utilized to find optimal hyperparameters for each machine learning model which helps in making the most accurate predictions.

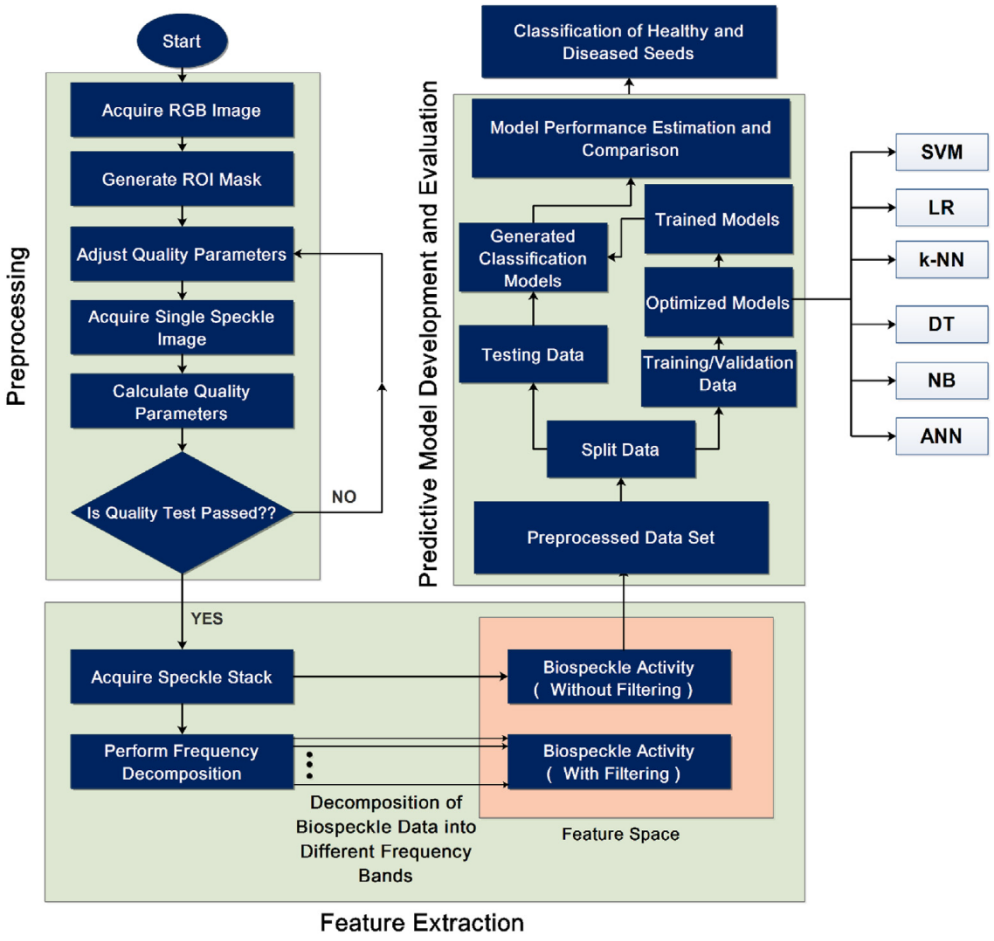


Fig. 3 – Flow chart for processing pipeline for data analysis including data pre-processing, feature extraction, model development and model evaluation process.

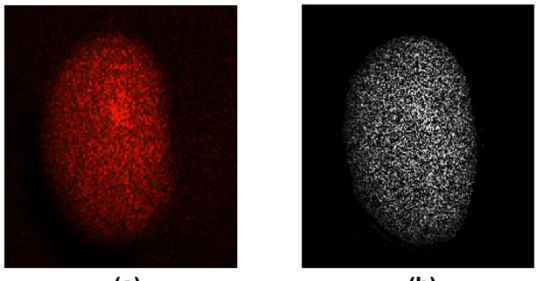


Fig. 4 – (a) Speckle image for soybean seed and (b) masked speckle image after background removal.

SVM is a supervised kernel-based discriminative binary classifier that analyses data points related to the classification and regression problems (Cortes & Vapnik, 1995). In this approach non-linear input data set is transformed into high dimensional linear feature space by using different kernels. In case of SVM, the penalty or regularisation (C) and kernel function are needed to be optimised for accurate prediction. Optimum value of C must be calculated such that the model is neither over-fitted nor under-fitted. In case of LR, L_1 and L_2 regularisation were used to compare the performance and the optimised function was compared with stochastic average gradient descent and Newton's method. k-NN is a non-parametric machine learning classifier used for classification and regression problems. For k-NN, the number of neighbours (k) and leaf size is the most important parameter to be optimised according to the specific task. DT is a supervised machine learning classifier that classify data items based on certain rules by using a tree-like structure. For DT optimum values of several parameters including the depth of tree, minimum samples number in each leaf node, minimum samples number to split a new node, maximum nodes, maximum samples, and features of the nodes were set for accurate prediction. NB is a probabilistic algorithm based on the probability theory and Bayes' theorem used to predict the intended classes. The applicability of Gaussian, Bernoulli, polynomial NB for classifying the seeds was tested. ANNs are a set of machine learning algorithms, inspired by the structure of biological neural networks of human brain. For the neural network classifier, we have optimised number of hidden layer and number of neurons in each layer. Multiple activation functions including sigmoid, Tanh, ReLU; and multiple optimiser functions including gradient decent, quasi-Newton, and Adam algorithm (Géron, 2019) were compared to get the best possible results. The rate of learning was set as adaptive mode by default with 100 epochs.

3.2.2. Model evaluation and performance estimation

All the acquired biospeckle activity information (without frequency filtering and with frequency filtering) associated with both the groups were pooled into a file and the observations were randomly selected to be used as training, validation, and testing data set. 60% of the data obtained from the experiments was used as a training set, 20% was used for parameter validation to optimise the trained model and 20% was used for testing the models. There was no intersection among these partitions. To ensure optimal utilisation of all the information residing in training data, the evaluation of trained/learned model is performed by using k-fold cross validation method (Rodriguez et al., 2009). In this method all data points are divided into defined number of subsets (k), out of which k−1 data points are put together to form a training set for the learning procedure of a model and the remaining one is used for evaluation. The approach was repeated for all the possible combinations of evaluation set and the average of all the possibilities was used to access the performance of generated model.

Figure 5 shows the confusion matrix for a two-class classifier. The confusion matrix (Rodriguez et al., 2009) comprises of the information related to actual and predicted classifications performed by a classification algorithm. To create the confusion matrix values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) were calculated. Finally, performance of generated models was evaluated by calculating several performance indicators including accuracy, specificity, precision, recall, and F1-score (Géron, 2019). Formulae for all the important parameters used in the study are given below:

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{6}$$

$$Specificity\ (\%) = \frac{TN}{TN + FP} \times 100 \tag{7}$$

$$Precision\ (\%) = \frac{TP}{TP + FP} \times 100 \tag{8}$$

$$Recall\ (\%) = \frac{TP}{TP + FN} \times 100 \tag{9}$$

$$F1 - score\ (\%) = \frac{2 \times Precision \times Recall}{Precision\ +\ Recall} \times 100 \tag{10}$$

The receiver operating characteristic (ROC) (Huang & Ling, 2005) curve is also one of the important tools used to evaluate the performance of the binary classifiers over a range of trade-offs between true positive rate (TPR) and false positive rate (FPR) (Géron, 2019). The values of TPR and FPR are also calculated by using the confusion matrix with the formulas given in Eqs. (11) and (12), respectively. To generate ROC curve TPR or recall is plotted against the FPR (where TPR is on the y-axis and FPR is on the x-axis). FPR is the ratio of negative instances that are incorrectly classified as positive. It is equal to one minus the true negative rate, which is the ratio of negative instances that are correctly classified as negative. Smaller values on the x-axis of the plot indicate lower false positives and higher true negatives. Larger values on the y-axis of the plot indicate higher true positives and lower false negatives.

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \tag{11}$$

$$False\ Positive\ Rate\ (FPR) = \frac{TN}{TN + FP} \tag{12}$$

Finally, area under the ROC curve (AU-ROC) is calculated from the formula given below:

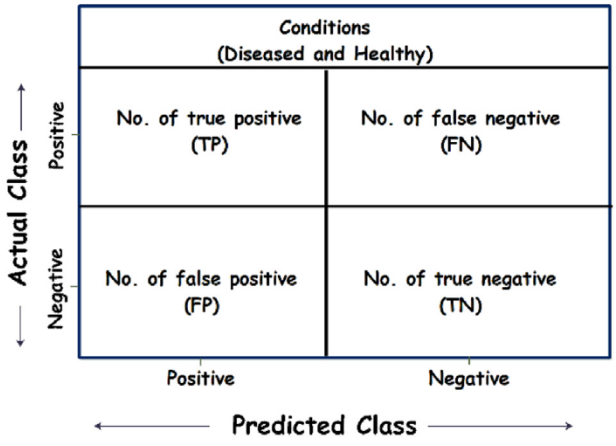


Fig. 5 – Representation of a two-class classifier confusion matrix.

Table 2 – Hyperparameters for machine learning models.

Classification Algorithm	Hyper-parameter Tuning
SVM	Regularisation C, Kernel Function (Linear and RBF), and gamma
LR	L1/L2 regularisation parameter
k-NN	Leaf size, K-neighbours
DT	Max depth, max leaf nodes, min leaf nodes, min sample leaf, min samples split
NB	Polynomial, Bernoulli, and Gaussian
ANN	Number of neurons and hidden layers, activation function, learning, number of epochs, and learning rate for the backpropagation algorithm, optimisation function

$$\text{Area under the ROC curve } (AU-ROC) = \frac{1}{2}(TPR + TNR) \quad (13)$$

3.3. Statistical analysis

Statistical analysis was conducted using IBM SPSS statistics software (version 16.0). All the data acquired from the measurements was subjected to analysis of variance (ANOVA) (Dudoit et al., 2003). Twenty measurements were made for each experiment and averaged to obtain an experimental unit. Every measurement corresponding to the healthy and diseased groups was compared to analyse the effect of seed-borne fungal pathogen statistically by considering 5% significance. Differences among the means of healthy and diseased groups were tested for significance by using Tukey's honest significant difference test (Tukey's HSD).

4. Results and discussion

4.1. Progression of seed-borne fungal pathogen with inoculums concentration and time

Growth of seed-borne fungal pathogen as a function of inoculum concentration and time was evaluated using biospeckle technique and benchmarked with reference to the percentage ACD. Biospeckle images of seeds were processed using AD based method; for interpretations, visual maps as well as numerical values of BA were evaluated. A substantial increase in the biospeckle activity with the inoculum concentration and the time after the inoculation (Table 3) was observed. Figure 6 shows the visual activity maps associated with healthy and diseased groups after 48 h of the treatments.

To get insight into the biospeckle behaviour of each sample, the normalised AD_{index} of both the groups, healthy (H) and diseased (D_2-D_6), were plotted against time and are shown in Fig. 7. The main objective of the work was to analyse the effect of seed-borne fungus pathogen of different inoculum concentrations on the biospeckle activity in early stages of infection. Hence, low level of inoculum concentration was applied to create infection in soybean seeds. Minimum potential inoculums require more than hundred fungal spores seed^{-1} for the establishment of infection which can produce typical symptom of disease (Rajput et al., 2020), otherwise seeds seem to be asymptomatic. However, concentration less than or equal to the specified number can also infect seeds and can result in severe losses after certain periods of time. Initially, just after the inoculation (0 h), no significant difference in AD_{index} values were obtained in samples of both the groups (diseased and healthy). This is due to the latency period of the fungi interaction with the seeds, as pathogen require certain time to grow in their hosts (Rennie and Cokerell, 2006). For the healthy group AD_{index} increased with time, however this change was relatively smaller as compared to the diseased group. Increase in the AD_{index} of healthy seeds was due to metabolic variation generated as a result of small water uptake during the treatment. This water uptake results in initiation of certain physiological and biochemical processes associated with the seed imbibition (Cardoso et al., 2011). As the concentration of inoculums and/or time duration of pathogen infection increased, continuous increase in AD_{index} was observed (Fig. 7). The highest change in AD_{index} detected was for seed treated for 48 h with 20000 spores seed^{-1} (D_6).

Table 3 summarises the results obtained for standard rating protocol (%ACD) as well as biospeckle techniques associated with healthy and diseased groups for different time durations. Percentage ACD was measured to compare and benchmark the results obtained by biospeckle technique. The descriptive statistics for the acquired data is also presented in Table 3. ANOVA was implemented on BA values for both, the healthy and the diseased groups to evaluate the influence of seed-borne fungus in the early stage of infection.

The healthy group does not present significant variation ($p < 0.05$) in AD_{index} with time after inoculation of seeds. As expected, time duration of the treatments and inoculum concentrations for diseased group has significant ($p < 0.05$) impact on the BA. Values of AD_{index} were significantly ($p < 0.05$) influenced by the interaction between the time duration and inoculum concentrations. The application of post-hoc test (Tukey's HSD) revealed that the BA of diseased seeds was significantly higher ($p < 0.05$) than that of the healthy group seeds. Seeds with inoculum concentration of 2 and 20 spores seed^{-1} appeared symptomless on inspection and percentage area covered by the disease was negligible. However, this small concentration can also colonise fungus with time progression, and pathogen can reside in seed coat with mixed infections as the seed coat is a common infection site for *C. truncatum* (Majumder et al., 2013). AD_{index} for small concentrations of inoculum responded well and increased significantly with time. It is clear, therefore, that biospeckle technique can detect the effects of fungal pathogen in its early stage of development. Further, even for higher inoculation concentration (2000 spores seed^{-1}) and after 36 h of inoculation, the area of seeds' surface covered by pathogen (Table 3) was significantly low (13.76%). For the maximum spore concentration (20,000 spores seed^{-1}), the seed shows visible signs of disease only after 24 h. Furthermore, even such high inoculum concentration was not detectable in early stage of pathogen development by using existing methods. The small biochemical and physiological changes initiated due to pathogen interaction were detectable using biospeckle technique, which were directly reflected on AD_{index}. The obtained results are very important as under controlled conditions, biospeckle analysis is not only able to distinguish between healthy and diseased seeds, but also detect the presence of pathogen in early stages of infection.

4.2. Frequency decomposition of biospeckle images for biological feature isolation in seed

As discussed, biospeckle laser technique allows optical monitoring of activities related to several physiological and biochemical phenomena, associated with various biological samples. In case of seeds, this activity is due to different vital metabolic processes namely reserve mobilisation, glyoxylate cycle, phytohormonal regulation, and respiration process in biological tissues (Ali and Elozeiri, 2017). Initially, seed imbibition triggers activation of various metabolic processes such

Table 3 – Biospeckle activity (BA) and percentage area covered by disease (%ACD) associated with differently treated seeds ((Data are means (±%SD)).

Time after inoculation (h)	Healthy Group		Diseased Group									
	H		D_2		D_3		D_4		D_5		D_6	
	AD_{index}	ACD (%)	AD_{index}	ACD (%)	AD_{index}	ACD (%)	AD_{index}	ACD (%)	AD_{index}	ACD (%)	AD_{index}	ACD (%)
0	320.21 (±0.32)^{a}	0	325.25 (±0.30)^{a}	0	322.59 (±0.42)^{a}	0	310.74 (±0.35)^{a}	0	324.00 (±0.21)^{a}	0	321.19 (±0.23)^{a}	0
12	343.72 (±0.22)^{ab}	0	453.56 (±0.34)^{ab}	0	496.33 (±0.31)^{b}	0	527.89 (±0.39)^{b}	0	585.44 (±0.41)^{b}	0	626.67 (±0.33)^{ab}	4.76
24	379.67 (±0.60)^{abc}	0	514.11 (±0.45)^{bc}	0	545.05 (±0.51)^{bc}	0	585.00 (±0.35)^{bc}	3.32	617.61 (±0.32)^{c}	6.46	650.83 (±0.42)^{bc}	12.49
36	404.89 (±0.47)^{bcd}	0	525.50 (±0.40)^{d}	0	558.06 (±0.37)^{cd}	2.10	593.67 (±0.44)^{d}	5.46	627.50 (±0.34)^{cd}	10.76	657.89 (±0.25)^{c}	23.73
48	409.11 (±0.70)^{cd}	0	536.78 (±0.29)^{cd}	0	569.61 (±0.42)^{d}	3.87	604.44 (±0.46)^{cd}	10.23	638.39 (±0.26)^{cd}	13.67	682.33 (±0.55)^{d}	31.89

The values are rounded off to the nearest integer up to two places of decimal for the given numerical data. The values superscripted with different letter in same column are significantly different at $p \leq 0.05$ (Tukey's Honest Significant Difference test).

as synthesis of hydrolytic enzymes which results in hydrolysis of reserve food into simple available form that can be further utilised by embryo to trigger germination processes. The phenomena can be activated just after enough water is available for imbibition of seeds. In the current study, experiments were planned in such a way that the seed could not be exposed to sufficient water intake, thereby prohibiting imbibition process. This also obstructs all the subsequent processes related to activation and germination of seeds (Ali and Elozeiri, 2017). Only small amount of water (10 μl) along with fungi spores was used for inoculation to introduce pathogen infection on seeds. However, small moisture content, if present, may also play an important role in seed physiology and can affect the seed metabolism. It may transport the nutrients which enhances the metabolic activity of the seed and can initiate the various biochemical processes related to hydration, development, or resting.

Hence, considering these factors, acquired raw biospeckle images were further divided into different frequency bands and analysed using AD based method. The main objective of the analysis was related to isolation of information about the BA generated due to interaction of seed with water. Figure 8 shows the change in normalised AD_{index} of different frequency bands (1.5–15 Hz) with varying time durations. AD_{index} corresponding to first frequency band (0–1.5 Hz) does not contain any detectable information; hence, this band is not shown in the graphical results. Initially, without frequency filtering, it was established that AD_{index} increases continuously as time duration of the pathogen interaction and/or inoculum concentration increases. Similar trends were observed for the frequency filtered data; AD_{index} for different frequency bands increased continuously as inoculum concentration increased, as shown in Fig. 8 (a)–(d). The obtained frequency signature also indicated that biospeckle analysis can distinguish between healthy and diseased seeds and can even examine the biological changes inside the seeds for different inoculum concentration levels associated with varying time durations.

A distinct change in the AD_{index} for different frequency regions was observed for the diseased group. As shown in Fig. 8, lower frequency bands (4.5 Hz and 6 Hz/3rd and 4th bands respectively) presented higher BA values for diseased seeds. Moreover, this change in AD_{index} is prominent and increased continuously with time progression and for higher inoculum concentrations as well. The observed biospeckle signature in these frequency bands can be attributed to the presence of fungal infection in the seed tissues. These speckle signatures for lower frequency bands were produced by the microscopic movements of micro-organism or other physical reasons including colonisation of fungi, inter and intra cellular infection, and mycelial growth. Conversely, continuous decrease in the BA of mid- and higher-frequency regions (7.5–10.5 and 12.0–15.0 Hz) was observed. BA in the mid-frequency regions (7.5–10.5 Hz) decreased linearly, however a rapid decrease was observed in high frequency regions. The observed decrease in BA at higher frequency bands (12.0–15.0 Hz) can be attributed to the effect of water activity, as low volume of water (approximate 10 μl) was provided, which does not allow the occurrence of seed imbibition or its related physiological activities. The interaction of water with

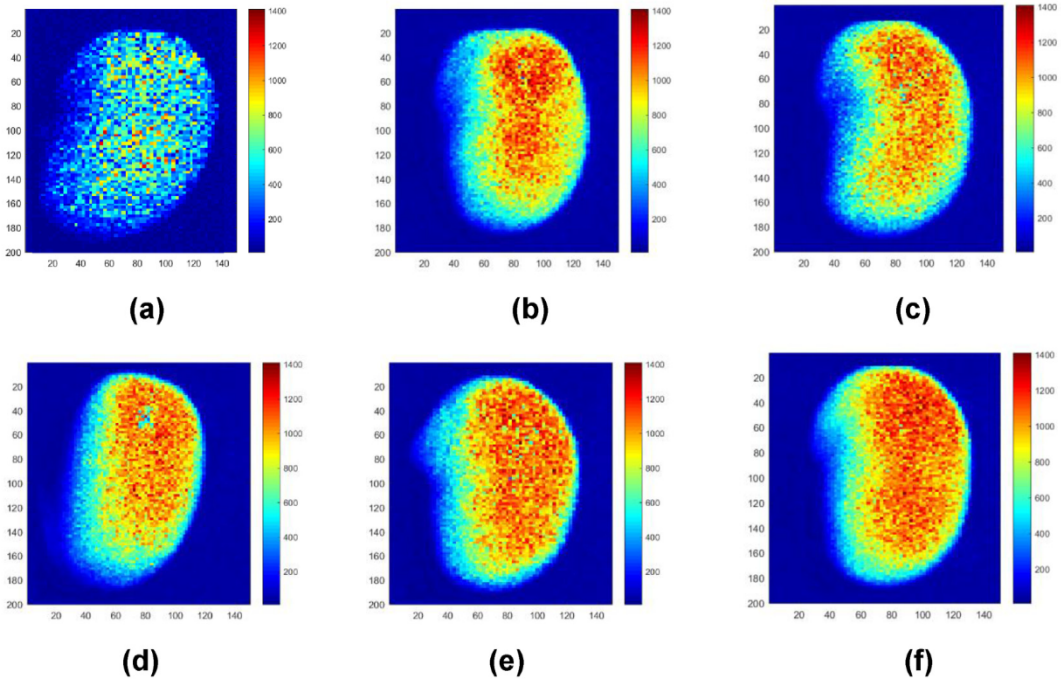


Fig. 6 – Activity map for differently treated seeds after 48 h for (a) H, (b) D_2, (c) D_3, (d) D_4, (e) D_5 and (f) D_6 treatments.

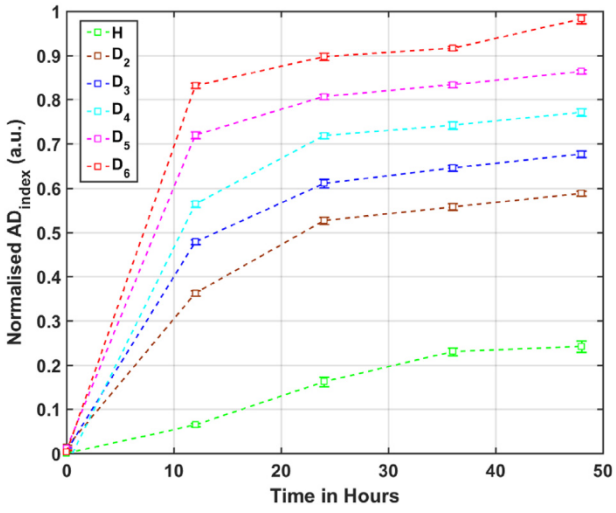


Fig. 7 – Normalised AD_{index} of healthy and diseased groups at different time intervals with varying inoculum concentration.

seed by using BA has also been reported in previous studies (Alves et al., 2013, Braga et al., 2005; Cardoso et al., 2011, Nobre et al., 2009, Sutton and Punja, 2017). The authors in these studies established that the BA occurring at higher frequencies is indicative of the seed's exposure time to water and morphological changes initiated due to germination progress. However, this was not the main physiological process during contamination; hence BA after removal of these frequency component was mainly due to the seed-borne fungi. Additionally, for mid-frequency range (7.5–10.5 Hz) the segmentation of activity suggests that the BA in this region decreases linearly. This change in biospeckle activity might be due to the mechanical process of water condensation on the seed surface (Cardoso et al., 2011). The reduction in the biospeckle activity is influenced by the seed traits related to dormancy and surface interaction of water in the seed. While considering the case of healthy seeds, almost flat frequency response was observed and most of the spectral components presented same index value. However, in this case also higher frequency resulted in small decrease in the AD_{index}, due to low moisture content. The current work presents a comprehensive procedure based on biospeckle analysis that can realize the isolation of biological phenomena in seeds, associated with seed-borne fungal pathogens by the means of frequency signatures.

4.3. *Automated classification of healthy and inoculated seeds with fungal pathogens at early stages of pathogen infection*

To differentiate between healthy and diseased seeds before specific symptoms became visible, several machine learning algorithms were implemented. Total 540 samples associated with healthy and diseased seeds were used to train and test the various machine learning models.

For SVM classifier, we have implemented both linear as well as RBF based kernels. The best results were observed when C is 1 and gamma is 0.1, where the model provided optimum performance (neither under-fitted nor over-fitted) with higher accuracy. We found that higher degrees for a polynomial based kernel required high computation time, while achieving approximately similar accuracy score.

For k-NN classifiers, the optimum performance of the classifier was obtained when the leaf size of kD-tree is 30. The optimal number of neighbouring points was set to 5 for getting the higher value of classification accuracy and other

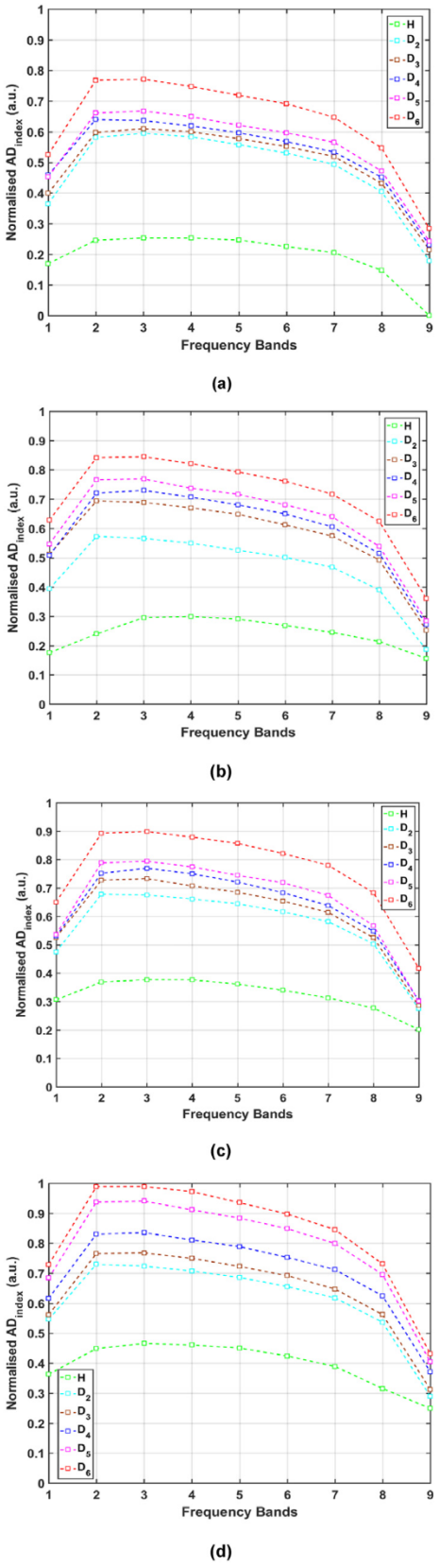


Fig. 8 – Normalised AD_{index} of healthy and diseased groups associated with different frequency bands for (a) 0 h, (b) 12 h, (c) 36 h, (d) 48 h, having varying inoculum concentration.

Table 4 – Confusion matrix obtained for different classifiers.

Classifier	Predicted/actual class	Healthy	Diseased
SVM	Healthy	102	12
	Diseased	6	312
LR	Healthy	80	34
	Diseased	25	293
k-NN	Healthy	106	16
	Diseased	1	309
DT	Healthy	102	12
	Diseased	1	317
NB	Healthy	76	38
	Diseased	59	259
ANN	Healthy	103	7
	Diseased	10	312

performance matrix. We have calculated the distance matrix (distance between each labelled data point in the data set to the point which is to be classified) by using Euclidean, Manhattan, Minkowski, and tangential distance matrix (Saini et al., 2013). We found that the Euclidean distance metrics obtain best results for the classification by using the uniform weights.

For LR classifiers, L1 and L2 regularisation method was utilised. However, L2 regularisation presented better results compared to L1 regularisation. For optimisation of loss function, stochastic gradient decent method was selected as this algorithm can attain better results for the given datasets.

For the decision tree classifier, the maximum depth of the tree was chosen to be 8. The maximum number of leaf nodes were set to be 3 for each tree, and minimum sample split was set to 2 (default value) for getting best possible results.

For the NB classifier, it was found that the Gaussian NB provided better classification accuracy as compared to the Bernoulli and polynomial NB classifiers.

For artificial neural network, we have used sequential model with 3 hidden layers containing 64 neuron each. For hidden layers, we have used ReLu as an optimisation function. For output layer, we have used Sigmoid as an optimisation function. A deeper neural network may result in better classification accuracy, but it will increase the training time and complexity of the model.

Assessment of classification models were performed by measuring several parameters with the help of confusion matrices given in Table 4. Table 5 provides the detailed classification results associated with the different models for classifying healthy and diseased seeds. Final analysis show that k-NN, DT and ANN models performed quite well among all classifiers with higher accuracy (96.79%, 96.94, and 96.37, respectively). Moreover, we have also obtained higher values of precision, recall, and F-1 scores for these classifiers using 10-fold cross validation method. Both k-NN and DT based learning models presented higher recall value (100%), which defined that these two classifiers correctly predicted entire diseased seeds from all the observations which actually belongs to the diseased group. Specificity of theses model was also very high 99.69 and 99.68% which imply that these models are also capable of accurately predicting healthy seeds which actually belongs to the healthy group. SVM classifier

Table 5 – Performance parameters obtained by different classifier for classifying diseased seeds.

Algorithms	SVM	LR	k-NN	DT	NB	ANN
Accuracy (%)	95.83	87.24	96.79	96.94	78.25	96.37
Specificity (%)	98.11	92.14	99.68	99.69	81.45	96.89
Precision (%)	96.40	90.17	95.24	96.32	87.46	91.15
Recall (%)	98.14	92.75	100.00	100.00	81.46	93.64
F1-score (%)	97.46	91.49	97.82	98.47	84.71	91.39
AU-ROC	0.92	0.81	0.98	0.95	0.74	0.98

came out to be next best possible predictive model with a classification accuracy of 95.83%, and higher precision, recall and F-1 scores. NB classifier proves to be the worst performing model with lowest value of classification accuracy (74.4%) specificity (81.45%), precision (87.46%), recall (81.46%) and F-1 score (84.71%).

As discussed in the above section, ROC curve provides visual insight into the performance of the predictive models at various FPR. Figure 9 (a)-(f) show the ROC curve for the comparison of binary predictive power of various machine learning models. The ROC curve models the performance of the classifiers over a range on TPR (proportion of actual diseased seeds that were correctly identified), to the FPR or 1 - specificity, (where specificity is the proportion of healthy seeds that were correctly identified). For an ROC curve the diagonal line reflects the performance of model which yields the positive or negative results dissimilar to the true

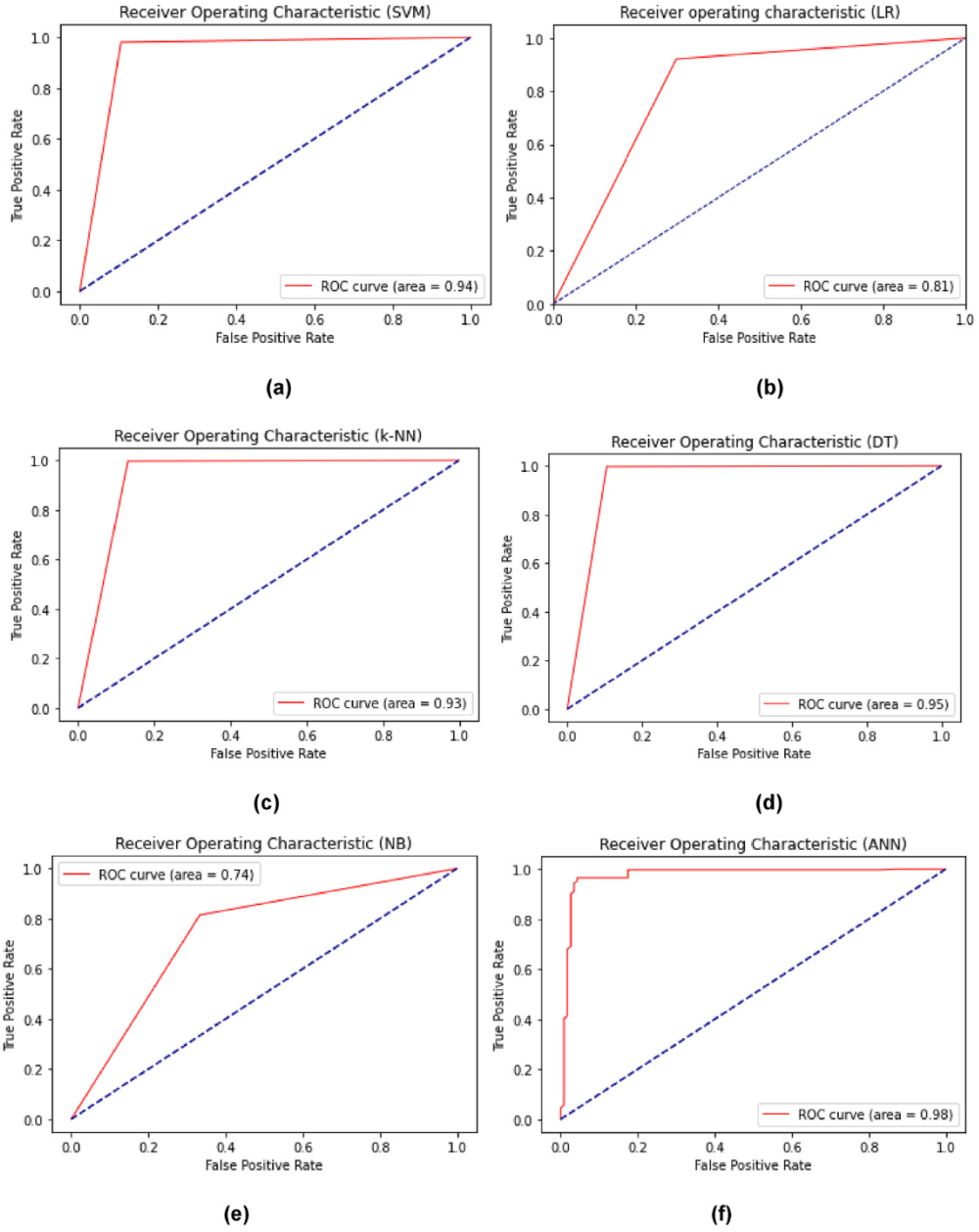


Fig. 9 – Receiver operating characteristics for all the machine learning models (a) SVM (b) LR (c) k-NN (d) DT (e) NB, and (f) ANN.

test results. Hence, the dotted blue line in Fig. 9 represents the ROC curve for a purely random classifier that does not provide any accurate prediction. This line is used to compare the performance of the several classifiers. For a good classifier the curve stays away from the diagonal line as much as possible (toward the top-left corner of the graph). One way to quantitively compare the performance of a classifiers by using ROC curve is to measure the AU-ROC. A perfect classifier without any misclassification will have AU-ROC score equal to 1, whereas for a random classifier the value of AU-ROC score is equal to 0.5. In the prediction of healthy and diseased seeds ANN, DT and k-NN based models possess top performance and AU-ROC score for the models are 0.98, 0.95 and 0.98 respectively. Values of AU-ROC scores for all the machine learning models for classifying healthy seeds from diseased one are given in Table 5. As already discussed, worst performing models in the class were LR and NB with low accuracy score. Similarly, for LR and NB classifier, the ROC curve is closer to the diagonal line and resulted in lower values of AU-ROC score (0.81 and 0.74 respectively). In conclusion, k-NN and DT classifiers can we used for a given classification task. However, ANN also gave better results, but the classifier is complex and required high computation time compared to other classifiers.

The obtained results establish that biospeckle analysis method in conjunction with machine learning based classifier can be a promising tool to classify the diseased seeds even in early stage of infection. We believe that the proposed technique can be very helpful to the plant pathologist, farmers and personnel associated with agricultural field, directly or indirectly, for arriving at decision regarding distinction between healthy and diseased seeds. By using such an efficient tool, accurate decisions for better plant protection and disease management can be made.

5. Conclusions

This study proved the feasibility of laser biospeckle technique in association with several machine learning algorithms for the early detection and automated classification of seeds infected with fungal pathogen. BA was determined by using AD based full field analysis technique and complemented by frequency filtering based approach to isolate the activity due to moisture inside the seeds. Significant increase in value ($p < 0.05$) of AD_{index} was obtained even when symptoms of infection were hardly observable using the benchmarked (percentage area covered by disease) technique. The performance of several machine learning classifiers was compared in terms of their feasibility to identify diseased seeds. k-NN, DT, and ANN proved to be best possible classifiers due to their better classification results and stronger robustness. Highlights of the proposed technique include its extreme simplicity, non-contact and non-invasive operations, high speed, low cost, requirement for only a small number of components, simple algorithms, and its ability to be commercialised. Moreover, machine learning based automatic detection procedure makes the method more suitable for real time field applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This publication is an outcome of the R&D work undertaken project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation, and Science and Engineering Research Board project grant (CRG/2018/002697).

Authors gratefully acknowledge Director, Indian Institute of Soybean Research (IISR) for the kind support towards this investigation. The authors are also grateful to the anonymous reviewers for several constructive suggestions. These have resulted in marked improvement of the content and presentation.

REFERENCES

Ali, A. S., & Elozeiri, A. A. (2017). Metabolic processes during seed germination. *Advances in Seed Biology*, 141–166.

Alves, J. A., Júnior, R. A. B., & Boas, E. V. D. B. V. (2013). Identification of respiration rate and water activity change in fresh-cut carrots using biospeckle laser and frequency approach. *Postharvest Biology and Technology*, *86*, 381–386. https://doi.org/10.1016/j.postharvbio.2013.07.030

Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 105635–105642. https://doi.org/10.1016/j.cmpb.2020.105635

Bock, C. H., Cook, A. Z., Parker, P. E., & Gottwald, T. R. (2009). Automated image analysis of the severity of foliar citrus canker symptoms. *Plant Disease*, *93*, 660–665. https://doi.org/10.1094/PDIS-93-6-0660

Braga, Roberto A., Jr., Rabelo, G. F., Granato, L. R., Santos, E. F., Machado, J. C., Arizaga, R., Rabal, H. J., & Trivi, M. (2005). Detection of fungi in beans by the laser biospeckle technique. *Biosystems Engineering*, *91*, 465–469. https://doi.org/10.1016/j.biosystemseng.2005.05.006

Braga, R. A., Jr., Horgan, G. W., Enes, A. M., Miron, D., Rabelo, G. F., & Barreto Filho, J. B. (2007). Biological feature isolation by wavelets in biospeckle laser images. *Computers and Electronics in Agriculture*, *58*, 123–132. https://doi.org/10.1016/j.compag.2007.03.009

Bramer, M. (2007). *Principles of data mining* (Vol. 180, pp. 31–38). London: Springer.

Cardoso, R. R., Costa, A. G., Nobre, C. M. B., & Braga, R. A., Jr. (2011). Frequency signature of water activity by biospeckle laser. *Optics Communications*, *284*, 2131–2136. https://doi.org/10.1016/j.optcom.2011.01.003

Carvalho, P. H., Barreto, J. B., Braga, R. A., Jr., & Rabelo, G. F. (2009). Motility parameters assessment of bovine frozen semen by biospeckle laser (BSL) system. *Biosystems Engineering*, *102*, 31–35. https://doi.org/10.1016/j.biosystemseng.2008.09.025

Chatterjee, A., Bhatia, V., & Prakash, S. (2017). Anti-spoof touchless 3D fingerprint recognition system using single shot fringe projection and biospeckle analysis. *Optics and Lasers in*

Engineering, 95, 1–7. https://doi.org/10.1016/j.optlaseng.2017.03.007

Chatterjee, A., Singh, P., Bhatia, V., & Prakash, S. (2018a). Application of random temporal indexing based laser biospeckle analysis for blood thrombocyte characterization. In *Presented at the international Conference on fiber Optics and photonics (PHOTONICS), IIT-Delhi*.

Chatterjee, A., Singh, P., Bhatia, V., & Prakash, S. (2018b). A low-cost optical sensor for secured antispoof touchless palm print biometry. *IEEE Sensors Letters*, 2, 1–4. https://doi.org/10.1109/LSENS.2018.2837879

Chatterjee, A., Singh, P., Bhatia, V., & Prakash, S. (2019). Ear biometrics recognition using laser biospeckled fringe projection profilometry. *Optics & Laser Technology*, 112, 368–378. https://doi.org/10.1016/j.optlastec.2018.11.043

Chatterjee, A., Singh, P., Bhatia, V., & Prakash, S. (2020). An efficient automated biospeckle indexing strategy using morphological and geo-statistical descriptors. *Optics and Lasers in Engineering*, 134, 1–12. https://doi.org/10.1016/j.optlaseng.2020.106217

Chen, F. L., & Li, F. C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, 37, 4902–4909. https://doi.org/10.1016/j.eswa.2009.12.025

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Dai Pra, A. L., Meschino, G. J., Guzmán, M. N., Scandurra, A. G., González, M. A., Weber, C., Trivi, M., Rabal, H., & Passoni, L. I. (2016). Dynamic speckle image segmentation using self-organizing maps. *Journal of Optics*, 18(85606), 1–11. https://doi.org/10.1088/2040-8978/18/8/085606

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18, 71–103.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). USA: O'Reilly Media.

Haykin, S. (1994). *Neural networks: A comprehensive foundation* (2nd ed., pp. 23–56). Singapore: Prentice Hall PTR.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17, 299–310. https://doi.org/10.1109/TKDE.2005.50

Kooij, H. M., Fokkink, R., Van Der Gucht, J., & Sprakel, J. (2016). Quantitative imaging of heterogeneous dynamics in drying and aging paints. *Scientific Reports*, 6, 34383–34391. https://doi.org/10.1038/srep34383, 10.

Kumar, R., & Gupta, A. (2020). *Seed-borne diseases of agricultural crops: Detection, diagnosis & management* (1st ed., pp. 107–142). Singapore: Springer Nature.

Larose, D. T. (2006). *Data mining methods and models* (1st ed.). Hoboken, NJ: Wiley-Interscience.

Madhusudhan, P., Sinha, P., Rajput, L. S., Bhattacharya, M., Sharma, T., Bhuvaneshwari, V., Gaikwad, K., Krishnan, S. G., & Singh, A. K. (2019). Effect of temperature on Pi54-mediated leaf blast resistance in rice. *World Journal of Microbiology and Biotechnology*, 35, 148. https://doi.org/10.1007/s11274-019-2724-8

Majumder, D., Rajesh, T., Suting, E. G., & Debbarma, A. (2013). Detection of seed borne pathogens in wheat: Recent trends. *Australian Journal of Crop Science*, 7, 500–507.

Mancini, V., & Romanazzi, G. (2014). Seed treatments to control seedborne fungal pathogens of vegetable crops. *Pest Management Science*, 70, 860–868. https://doi.org/10.1002/ps.3693

Moreira, J., Cardoso, R. R., & Braga, R. A. (2014). Quality test protocol to dynamic laser speckle analysis. *Optics and Lasers in Engineering*, 61, 8–13. https://doi.org/10.1016/j.optlaseng.2014.04.005

Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. *Procedia Technology*, 4, 119–128. https://doi.org/10.1016/j.protcy.2012.05.017

Nataraj, V., Maranna, S., Kumawat, G., Gupta, S., Rajput, L. S., Kumar, S., Sharma, A. N., & Bhatia, V. S. (2020). Genetic inheritance and identification of germplasm sources for anthracnose resistance in soybean [*Glycine max (L.) Merr.*]. *Genetic* Resources And Crop Evolution, 1–8. https://doi.org/10.1111%2Fmpp.13036.

Nobre, C. M. B., Braga, R. A., Jr., Costa, A. G., Cardoso, R. R., Da Silva, W. S., & Sáfadi, T. (2009). Biospeckle laser spectral analysis under inertia moment, entropy and cross-spectrum methods. *Optics Communications*, 282, 2236–2242. https://doi.org/10.1016/j.optcom.2009.02.061

Polat, K., & Güneş, S. (2007). Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Applied Mathematics and Computation*, 187, 1017–1026. https://doi.org/10.1016/j.amc.2006.09.022

Rabelo, G. F., Enes, A. M., Junior, R. A. B., & Dal Fabbro, I. M. (2011). Frequency response of biospeckle laser images of bean seeds contaminated by fungi. *Biosystems Engineering*, 110, 297–301. https://doi.org/10.1016/j.biosystemseng.2011.09.002

Rajput, L. S., Harlapur, S. I., Venkatesh, I., Aggarwal, S. K., & Choudhary, M. (2016). In-vitro study of fungicides and an antibiotic against Rhizoctonia solani f. sp. sasakii causing banded leaf and sheath blight of maize. *International Journal of Agricultural Science*, 54, 2846–2848.

Rajput, L. S., Madhusudhan, P., & Sinha, P. (2020). Seed-borne diseases of agricultural crops: Detection, diagnosis & management. In R. Kumar, & A. Gupta (Eds.), *Epidemiology of seed-borne diseases*. Singapore: Springer.

Rajput, L. S., Sharma, T., Madhusudhan, P., & Sinha, P. (2017). Effect of temperature on growth and sporulation of rice leaf blast pathogen Magnaportheoryzae. *International Journal of Current Microbiology and Applied Sciences*, 6, 394–401. https://doi.org/10.20546/ijcmas.2017.603.045

Rennie, W. J., & Cokerell, V. (2006). Seedborne diseases. In B. M. Cooked, G. Jones, & B. Kaye (Eds.), *Epidemiology of plant diseases* (pp. 357–372). Springer (Chapter 13).

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 569–575. https://doi.org/10.1109/TPAMI.2009.187

Rumpf, T., Mahlein, A. K., Steiner, U., Oerke, E. C., Dehne, H. W., & Plümer, L. (2010). Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 74, 91–99. https://doi.org/10.1016/j.compag.2010.06.009

Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of Advanced Research*, 4, 331–344. https://doi.org/10.1016/j.jare.2012.05.007

Schaad, N. W., Frederick, R. D., Shaw, Schneider, W. L., Robert, H., Michael, D. P., & Luster, D. G. (2003). Advances in molecular-based diagnostics in meeting crop biosecurity and phytosanitary issues. *Annual Review of Phytopathology*, 41, 305–324. https://doi.org/10.1146/annurev.phyto.41.052002.095435

Schütz, G., Haltrich, D., & Atanasova, L. (2020). Influence of spore morphology on spectrophotometric quantification of Trichoderma inocula. *Biotechniques*, 68, 279–282. https://doi.org/10.2144/btn-2019-0152

Sendra, G. H., Arizaga, R., Rabal, H., & Trivi, M. (2005). Decomposition of biospeckle images in temporary spectral bands. *Optics Letters, 30*, 1641–1643. https://doi.org/10.1364/OL.30.001641

Singh, P., Chatterjee, A., Bhatia, V., & Prakash, S. (2018). *A mobile phone based low cost biospeckle analysis tool for real time applications. presented at the* International Conference on Fiber Optics and Photonics (PHOTONICS), *IIT-Delhi*.

Singh, P., Chatterjee, A., Bhatia, V., & Prakash, S. (2020a). Discrete cosine transform based processing framework for indexing, decomposition and compression of biospeckle data. *Laser Physics, 30*, 1–13. https://doi.org/10.1088/1555-6611/ab9021

Singh, P., Chatterjee, A., Bhatia, V., & Prakash, S. (2020b). Application of laser biospeckle analysis for assessment of seed priming treatments. *Computers and Electronics in Agriculture, 169*. https://doi.org/10.1016/j.compag.2020.105212, 105212-1-12.

Sutton, D. B., & Punja, Z. K. (2017). Investigating biospeckle laser analysis as a diagnostic method to assess sprouting damage in wheat seeds. *Computers and Electronics in Agriculture, 141*, 238–247. https://doi.org/10.1016/j.compag.2017.07.027

Tao, F., Yao, H., Hruska, Z., Burger, L. W., Rajasekaran, K., & Bhatnagar, D. (2018). Recent development of optical methods in rapid and non-destructive detection of aflatoxin and fungal contamination in agricultural products. *Trends in Analytical Chemistry, 100*, 65–81. https://doi.org/10.1016/j.trac.2017.12.017

Zdunek, A., Adamiak, A., Pieczywek, P. M., & Kurenda, A. (2014). The biospeckle method for the investigation of agricultural crops: A review. *Optics and Lasers in Engineering, 52*, 276–285. https://doi.org/10.1016/j.optlaseng.2013.06.017

Zdunek, A., & Cybulska, J. (2011). Relation of biospeckle activity with quality attributes of apples. *Sensors, 11*, 6317–6327. https://doi.org/10.3390/s110606317