International Journal of Biometeorology
https://doi.org/10.1007/s00484-024-02707-4

ORIGINAL PAPER

Comparative analysis of different Karnal bunt disease prediction models developed by machine learning techniques for Punjab conditions

Shubham Anand[1] · Sarabjot Kaur Sandhu[1] · Barun Biswas[2] · Ritu Bala[3]

Received: 30 October 2023 / Accepted: 18 May 2024

Abstract

Timely prediction of pathogen is important key factor to reduce the quality and yield losses. Wheat is major crop in northern part of India. In Punjab, wheat face challenge by different diseases so the study was conducted for two locations viz. Ludhiana and Bathinda. The information regarding the occurrence of Karnal bunt in 12 consecutive crop seasons (from 2009-10 to 2020-21) in Ludhiana district and in 9 crop seasons (from 2010-11 to 2018-19) in Bathinda district, was collected from the Wheat Section of the Department of Plant Breeding and Genetics at Punjab Agricultural University (PAU), located in Ludhiana. The study aims to investigate the adequacy of various methods of machine learning for prediction of Karnal bunt using meteorological data for different time period viz. February, March, 15 February to 15 March and overall period obtained from Department of Climate Change and Agricultural Meteorology, PAU, Ludhiana. The most intriguing outcome is that for each period, different disease prediction models performed well. The random forest regression (RF) for February month, support vector regression (SVR) for March month, SVR and BLASSO for 15 February to 15 March period and random forest for overall period surpassed the performance than other models. The Taylor diagram was created to assess the effectiveness of intricate models by comparing various metrics such as root mean square error (RMSE), root relative square error (RRSE), correlation coefficient (r), relative mean absolute error (MAE), modified D-index, and modified NSE. It allows for a comprehensive evaluation of these models' performance.

Keywords Karnal bunt · Wheat · Machine learning methods · Meteorological parameters · Disease prediction models

Introduction

The rapid population growth and unpredictable climate changes present significant challenges to the agricultural sector, particularly in terms of ensuring food security, productivity, and sustainability. Climate change has emerged as a critical concern affecting a country's food security, leading to extreme weather events. With temperatures projected to rise by 1-2.5°C by 2030, crop yields could be substantially impacted due to changes in photosynthesis, increased plant respiration, and alterations in disease incidence and pest populations (Bhanumathi et al. 2019).

Crop diseases are heavily influenced by weather conditions. The disease triangle, a conceptual model, outlines the fundamental factors responsible for causing disease. It explains that diseases occur when a virulent pathogen interacts with a susceptible host organism under favorable environmental conditions. The disease triangle was first depicted by Stevens in 1960 and was later revisited by Francl in 2001, who expanded the concept to include additional parameters such as humans, vectors, and time. Numerous researchers have dedicated considerable efforts to studying how various weather parameters interact and indirectly impact the development of plant disease outbreaks. These studies underscore the importance of incorporating multiple climate change parameters into models addressing this issue. Newbery et al. (2016) introduced a graphical scheme to facilitate a more concise understanding of how climate, crop growth, and disease models

✉ Shubham Anand
shubham-1904001@pau.edu

[1] Department of Climate Change & Agricultural Meteorology, PAU, Ludhiana, India

[2] Regional Research Station, Gurdaspur, India

[3] Department of Plant Breeding and Genetics, PAU, Ludhiana, India

Published online: 28 May 2024

can be integrated to project crop growth stages and disease incidence/severity under different climate change scenarios.

The issue of effective plant disease protection is closely linked to the challenges of sustainable agriculture and climate change. Climate change has the potential to impact pathogen development stages and rates, as well as alter host plant resistance, resulting in physiological changes in the interactions between hosts and pathogens. These changes can have significant implications for the occurrence and management of plant diseases (Garret et al. 2006). Several minor diseases have reappeared during different crop seasons. For instance, Karnal bunt, a significant wheat disease in Punjab, exhibited an upward trend in both severity and prevalence between 2012-13 and 2014-15 (Kaur et al. 2018). Despite being considered a minor disease, Karnal bunt resurfaced during the 2014-15 crop season due to the presence of favorable weather conditions (Sharma et al. 2012). Smiley (1996) emphasized the significance of specific climatic conditions, including appropriate rainfall and associated humidity levels, which play a crucial role in teliospore germination, secondary sporidial multiplication, penetration, and infection. These events need to be synchronized with the susceptible period, typically spanning 3 to 4 weeks leading up to wheat anthesis. The defined suitable rain and humidity events involve measurable rainfall (> 3 mm) occurring on two or more successive days, with at least 10 mm collected within a 2-day interval, and an average daily relative humidity exceeding 70% on both rainy days. To summarize, the climatic conditions during the susceptible period before wheat anthesis, which facilitate the survival, establishment, and spread of *Tilletia indica* sporidia, include optimum maximum temperatures ranging from 16 to 23°C, optimum minimum temperatures ranging from 7 to 11°C, high average daily humidity (> 70%) or relative humidity exceeding 48% at 3 pm, and measurable rainfall on multiple successive days (Smiley 1996). In context of Punjab, the favorable conditions for Karnal bunt were determined to be a maximum temperature ranging from 25 to 31°C in March, a minimum temperature ranging from 8.5 to 11.0°C in February, morning and evening relative humidity ranging from 85 to 95% and 40 to 60%, respectively, in March, and sunshine hours of 5.5 to 9.0 hours per day, along with rainfall exceeding 25 mm during mid-February to mid-March (Sandhu et al. 2022).

The liberalization of trade has facilitated the global spread of diseases, leading to the emergence of new diseases in previously unaffected regions where there may be a lack of local expertise to deal with them. Inadequate pesticide usage can also lead to the development of long-term resistance in pathogens, making it difficult for host plants to defend themselves. As a result, timely detection of diseases in plants poses a significant challenge for farmers (Anonymous 2019).

One potential solution to address this challenge is the development of prediction models based on the relationship between prevailing weather conditions and disease severity. By studying the complex interplay between plants, pathogens, and the environment, such models can aid in guiding management decisions. However, the complexity of many plant diseases and their dependence on mathematical or statistical forecasting models can be limiting. Although numerous laboratory studies have been conducted to understand the impact of environmental conditions on the survival and growth of T. indica (Smilanick et al. 1989), and several models have been created to simulate meteorological factors relevant to the establishment and spread of the disease (Jhorar et al. 1992, Mavi et al. 1992, Kaur et al. 2000, Sandhu et al. 2022), the task remains challenging due to the intricate nature of the disease processes involved. Several attempts have also been made to model KB forecasting in Indian conditions (Srinivasan 1980, Jhorar et al. 1992, Mavi et al. 1992; Singh et al. 1996) but all the models could not be validated in Punjab, India (Kaur et al. 2006). Biswas et al. (2013) carried out bivariate probability density analysis to develop a predictive regression model for inoculum load that was successfully validated and could be used for prediction of sporidial activity in field. Much of the recent progress has come from advances in computing and storage capabilities that are expected to improve complex computing systems that can learn to mimic humans and perform specific tasks. Biswas et al. (2013) carried out bivariate probability density analysis to develop a predictive regression model for inoculum load that was successfully validated and could be used for prediction of sporidial activity in field.

Much of the recent progress has come from advances in computing and storage capabilities that are expected to improve complex computing systems that can learn to mimic humans and perform specific tasks. Artificial Intelligence (AI) highlights the potential usefulness of pattern and trend detection in large amounts of data using pertinent mathematical algorithms and the objective of solving a particular task (Winston 1992). The task can be generic, such as computer vision, natural language processing (NLP), predictive modelling, or specific and related to a specific area that would otherwise require an expert in the field. AI includes field of Machine Learning (ML) and Deep Learning (DL). While, AI is the general term used to categorize any task that allows a machine or system to mimic human behaviour and intelligence, machine learning and deep learning are the specific methods used to do so. Machine learning uses algorithms that learn from data, make generalizations, and create rules that enable prediction of one or more target variables on the basis of set of input variables (Goodfellow et al. 2016). Fortunately, machine learning not only helps in understanding and developing new models but also accounts for understanding highly complex relationships to define with mathematical models (Fu et al. 2018). Thus, keeping this aspect in view this article proposes improved ML algorithms that use specialized ensemble methods such as artificial neural networks (ANN), efficient neural network (ENET), k-nearest neighbour (kNN), least absolute shrinkage and selection operator (LASSO), Bayesian least absolute shrinkage and selection operator (BLASSO), support vector regression (SVR), ridge regression (RIDGE), Bayesian ridge (BRIDGE), multiple linear regression (MLR), principal component regression (PCR), and random forest (RF) for developing prediction models for Karnal bunt disease.

Data methodology

Study area

The study was conducted at two locations viz. Ludhiana (latitude 30^{o}54', longitude N 75^{o}48'E and at an altitude of 247 meter above mean sea level) and Bathinda (latitude 30°58'N, longitude 74°18'E, altitude 211 meter above mean sea level. Ludhiana is located in the central plain region of Punjab with general climatic conditions classified as subtropical and semi-arid while Bathinda region falls in western zone and its climate is classified as semi-arid. Annual normal rainfall levels of Ludhiana and Bathinda are 760 mm and 436 mm, respectively. In Ludhiana, the summer temperature exceeds 40°C with dry summer spell while the lowest temperature may be near 0°C during winter season. In Bathinda, dust storms are common in May-June when the mercury touches 47°C and frosty nights associated with chilled winds are common when night temperature touches 0°C during December-January.

Disease data collection

The data on Karnal bunt incidence for 12 crop seasons (from 2009-10 to 2020-21) in Ludhiana district and 9 crop seasons (from 2010-11 to 2018-19) in Bathinda district was obtained from the Wheat Section of the Department of Plant Breeding and Genetics at PAU, Ludhiana (Fig. 1). To gather the Kb incidence data, wheat grain samples were collected from various grain markets in both districts. Approximately 15-30 samples of grains, each weighing between 500g to 1000g, were randomly collected from different wheat heaps and placed in paper bags.

$$\text{Disease incidence}(\%) = \frac{\text{No. of infected grains}}{\text{Total no. of grains examined}} \times 100 \quad (1)$$

Meteorological data collection

The meteorological data for the respective districts (Figs. 2, 3 and 4) was collected from the Department of Climate Change and Agricultural Meteorology at PAU, Ludhiana.

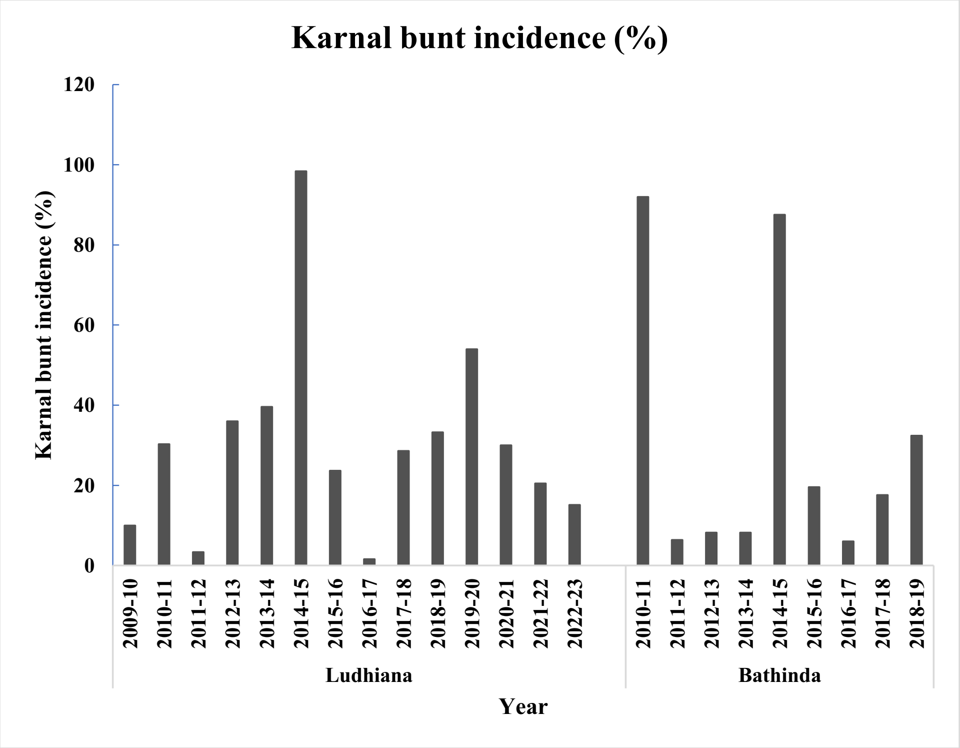

Fig. 1 Karnal bunt incidence data of Ludhiana and Bathinda

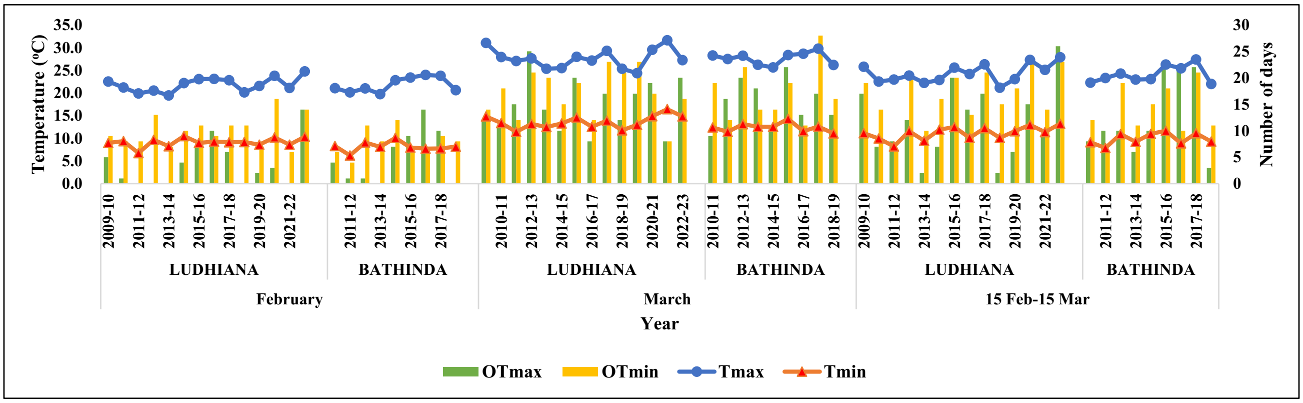

Fig. 2 Maximum and minimum temperatures of February and March months of Ludhiana and Bathinda

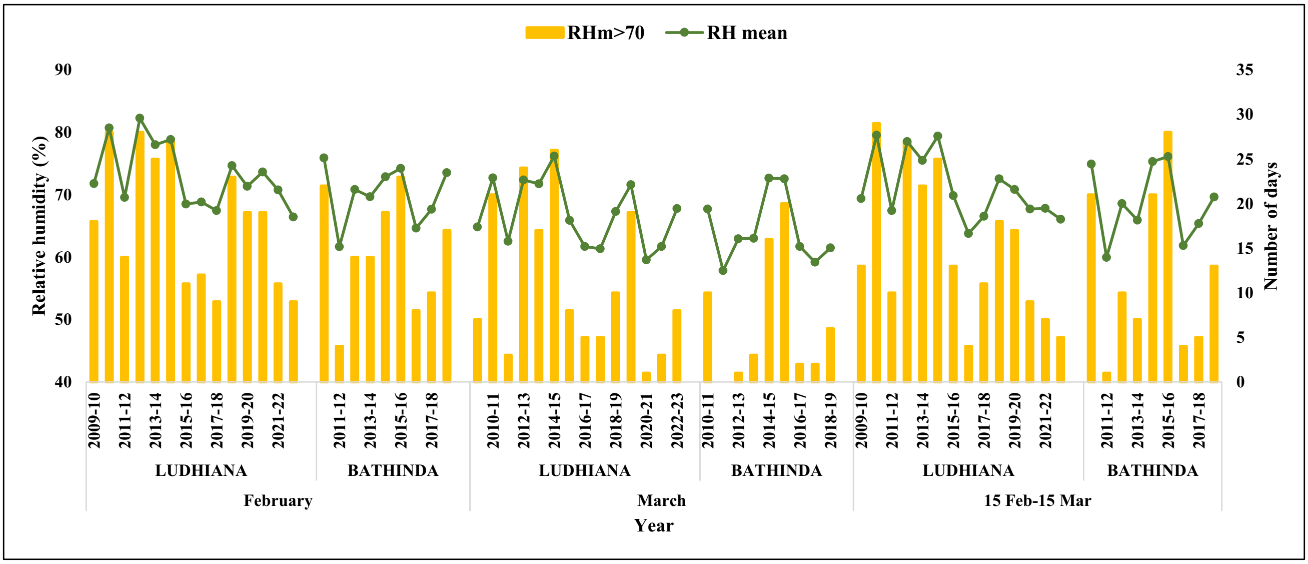

Fig. 3 Mean relative humidity of February and March months of Ludhiana and Bathinda

This weather data included maximum and minimum temperatures (Tmax and Tmin), mean relative humidity (RHme), rainfall (RF), and the number of rainy days (RD) for the months of February and March. These specific months were chosen as they correspond to the anthesis and ear formation stages of wheat, which are considered the most vulnerable stages for the development of Karnal bunt.To begin the analysis, descriptive statistics were applied to the meteorological data. Subsequently, a Humid-thermal index (HTI) was developed to forecast the suitability for disease establishment and spread. The HTI was calculated using the following formula:

$$\text{Humid-thermal index} = \frac{\text{Evening relative humidity}}{\text{Maximum temperature}} \tag{2}$$

Results of HTI are interpreted as per Table 1 in Fig. 5.

Potential predictor attributes

Eleven attributes were chosen as possible predictor variables (as shown in Table 2), and many of these attributes have been recognized as significant factors in previous studies regarding the development of Karnal bunt disease.

Machine learning regression models

The collected dataset was split into training and testing sections. The 70 per cent of the total dataset was used as training dataset while the remaining 30 per cent dataset was used as testing data. Machine learning regression models were applied to the dataset to train the model to predict disease. The process of modelling is shown in Fig. 6. These models include artificial neural networks

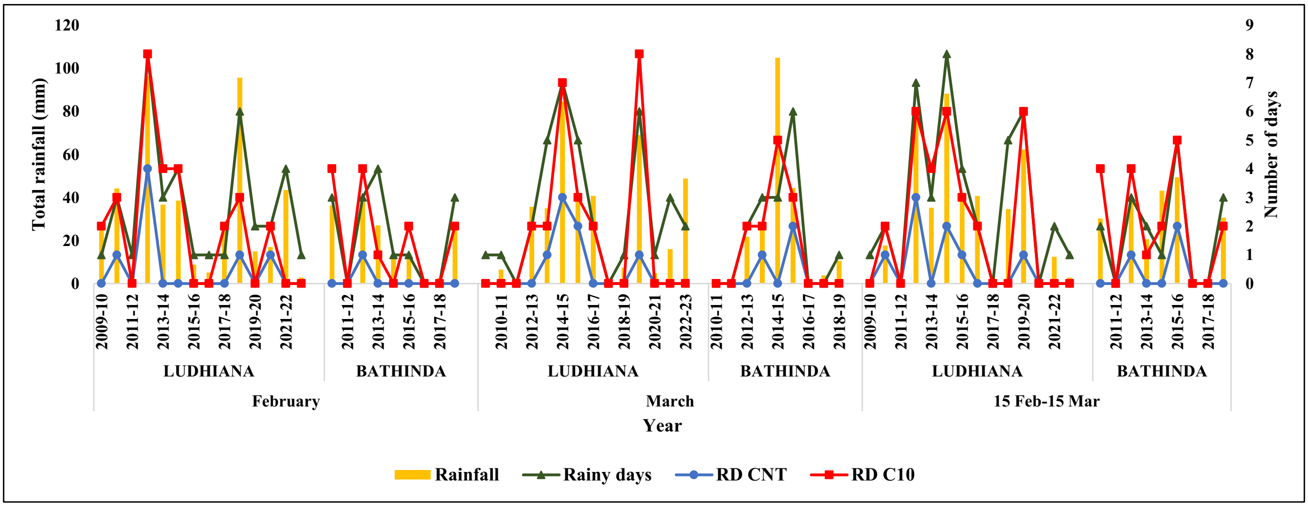

Fig. 4 Rainfall and rainy days of February and March months of Ludhiana and Bathinda

Table 1 Forecasting suitability for disease establishment and spread (source: Jhorar et al. 1992)

HTI	Forecasting suitability
< 2.2	Too hot or too dry
>2.2 and < 3.3	Suitable for Karnal bunt establishment and spread
> 3.3	Too cold or too wet

(ANN), efficient neural network (ENET), k-nearest neighbour (kNN), least absolute shrinkage and selection operator (LASSO), Bayesian least absolute shrinkage and selection operator (BLASSO), support vector regression (SVR), ridge regression (RIDGE), Bayesian ridge (BRIDGE), multiple linear regression (MLR), principal component regression (PCR), and random forest (RF).

Model accuracy terms/indices

Six of the most common accuracy metrics of regression models were used: root mean square error (RMSE), root relative square error (RRSE), correlation coefficient (r), the relative mean absolute error (MAE), modified D-index and modified NSE. Table 3 shows regarding estimation of these matrics.

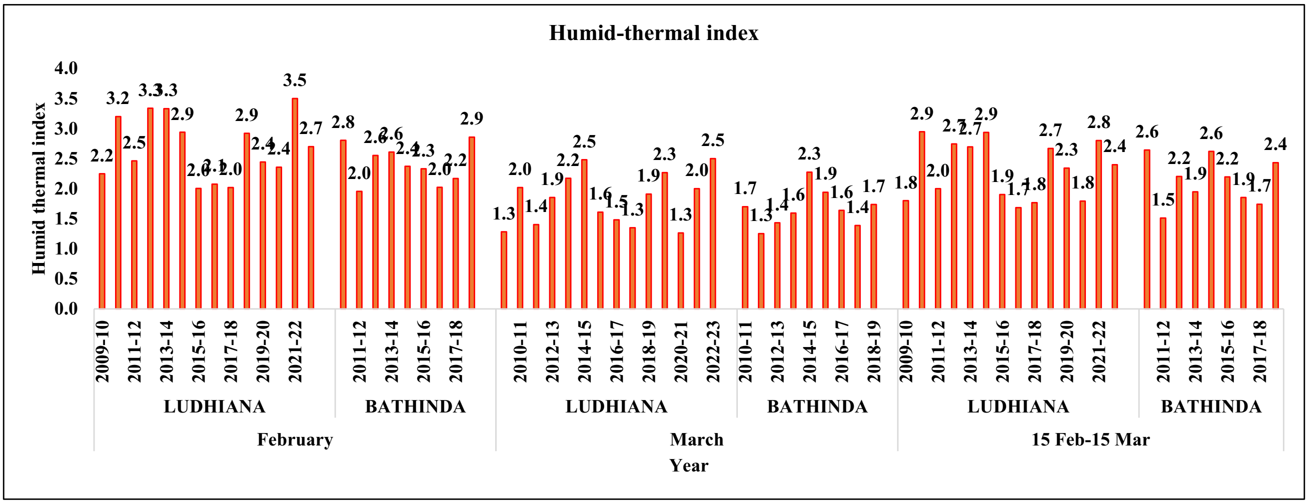

Fig. 5 Humid thermal index of February and March months of Ludhiana and Bathinda

Table 2 Potential predictor attributes

Attribute code name	Attribute name / description
Tmax	Maximum temperature
Tmin	Minimum temperature
OTmax	Number of weeks when optimum maximum temperature for Karnal bunt disease development occurred
OTmin	Number of weeks when optimum minimum temperatures for Karnal bunt disease development occurred
RHm	Mean relative humidity
RHm>70	Number of weeks when mean relative humidity was greater than 70 per cent
RF	Rainfall
RD	Rainy days
RD CNT	Number of weeks with continuous rainy days
RD C10	Number of days when at least 10 mm rainfall was recorded
HTR	Humid thermal ratio

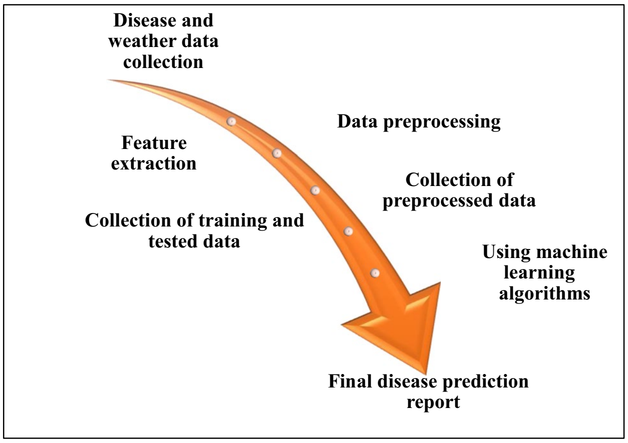
Disease and weather data collection
Data preprocessing
Feature extraction
Collection of preprocessed data
Collection of training and tested data
Using machine learning algorithms
Final disease prediction report

Fig. 6 Process of modelling

Results and discussion

Descriptive statistics

The descriptive statistics of studied weather parameters is presented in Tables 4, 5, and 6. In these tables, ranges of different meteorological parameters along with mean and standard deviation are presented for the periods under study. The mean maximum temperature during March month was mostly higher (27.34°C) than February (21.68°C) month and 15 February-15 March period (23.98°C). The mean number of days when optimum maximum temperature prevailed was higher (15.00) during March as compared to 15 February-15 March period (10.58) and February (3.17).

Table 3 Model accuracy terms/ indices

Metric	Description	Formula	Reference
Root mean square error (RMSE)	The RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers. The RMSE values should be as low as possible for a better-performing model.	$\sqrt{\frac{\sum_{i=1}^{n}\left(yi+\hat{y}_{\overline{y}}\right)^{2}}{n}}$	(Han & Kamber, 2006)
Root relative square error (RRSE %)	RRSE, which compares the model prediction against the mean. For this metric, a value below 100% indicates a better performance than the average.	$\sqrt{\frac{\sum_{i=1}^{n}\left(yi+\hat{y}_{i}\right)^{2}}{\sum_{i=1}^{n}\left(yi-\overline{y}\right)^{2}}\times 100}$	Gonzalez-sanchez et al. (2014)
Mean absolute error (MAE (%)	MAE is the average of differences in estimations (in physical units)	$\frac{\sum_{i=1}^{n}\left\|y_{i}-\hat{y}_{i}\right\|}{n\overline{y}}\cdot 100$	Gonzalez-sanchez et al. (2014)
Correlation coefficient (r)	Correlation coefficient (r) measures the linear relationship between regression model predictions and the real values.	$\frac{\sum_{i=1}^{n}\left(yi-\overline{y}\right)\left(\hat{y}i-\overline{y}\right)^{2}}{\sqrt{\sum_{i=1}^{n}\left(y_{i}-\overline{y}\right)^{2}}\sqrt{\sum_{i=1}^{n}\left(\hat{y}_{i}-\overline{y}\right)^{2}}}$	Gonzalez-sanchez et al. (2014)
Modified Index of aggrement (d)	The index of agreement represents the ratio of the mean square error and the potential error. The range of d is similar to that of r2 and lies between 0 (no correlation) and 1 (perfect fit)	$d_{j}=1-\frac{\sum_{i=1}^{n}\left\|O_{i}-P_{i}\right\|^{j}}{\sum_{i=1}^{n}\left(\left\|P_{i}-\overline{O}\right\|+\left\|o_{i}-\overline{o}\right\|\right)}$	(Willmot 1984)
Modified Nash–Sutcliffe efficiency (E)	The efficiency E proposed by Nash and Sutcliffe (1970) is defined as one minus the sum of the absolute squared differences between the predicted and observed values normalized by the variance of the observed values during the period under investigation.	$E_{j}=1-\frac{\sum_{i=1}^{n}\left\|Oi-P_{i}\right\|^{j}}{\sum_{i=1}^{n}\left\|O_{i}-\overline{O}\right\|^{j}}$	Nash and Sutcliffe (1970)

Where, y = real value, y^ = predicted value, i = observation, y¯, y^¯ = means, O observed and P predicted values

Table 4 Descriptive statistics of studied parameters for Karnal bunt of wheat for February month

Parameter	Tmax	Tmin	OTmax	OTmin	RHm	RHm>70	RF	RD	RD CNT	RD C10	HTR
Mean	21.68	9.07	3.17	10.00	73.80	19.42	34.51	2.75	0.58	2.33	2.61
Standard Error	0.41	0.28	0.95	0.76	1.47	1.98	9.10	0.65	0.34	0.68	0.15
Median	21.87	9.17	2.50	9.50	72.71	19.00	26.00	2.00	0.00	2.00	2.45
Standard Deviation	1.44	0.96	3.30	2.63	5.08	6.84	31.52	2.26	1.16	2.35	0.51
Sample Variance	2.06	0.93	10.88	6.91	25.81	46.81	993.32	5.11	1.36	5.52	0.26
Kurtosis	-1.37	2.64	-0.17	1.17	-1.26	-1.45	0.89	1.52	7.73	2.00	-1.56
Skewness	-0.21	-1.10	0.81	1.05	0.43	-0.13	1.33	1.45	2.66	1.19	0.30
Range	4.33	3.77	10.00	9.00	14.79	19.00	91.80	7.00	4.00	8.00	1.33
Minimum	19.45	6.73	0.00	7.00	67.46	9.00	4.60	1.00	0.00	0.00	2.00
Maximum	23.77	10.50	10.00	16.00	82.25	28.00	96.40	8.00	4.00	8.00	3.34

Table 5 Descriptive statistics of studied parameters for Karnal bunt of wheat for March month

Parameter	Tmax	Tmin	OTmax	OTmin	RHm	RHm>70	RF	RD	RD CNT	RD C10	HTR
Mean	27.34	13.27	15.00	17.92	67.31	12.17	27.17	2.58	0.58	2.00	1.76
Standard Error	0.57	0.33	1.40	1.14	1.58	2.52	8.27	0.71	0.29	0.81	0.12
Median	27.40	13.23	14.50	18.50	66.60	9.00	21.20	1.50	0.00	1.00	1.73
Standard Deviation	1.98	1.13	4.84	3.94	5.48	8.74	28.65	2.47	1.00	2.80	0.42
Sample Variance	3.92	1.28	23.45	15.54	30.01	76.33	821.00	6.08	0.99	7.82	0.17
Kurtosis	-0.56	-0.82	0.08	-1.23	-1.42	-1.49	-0.26	-1.06	2.23	1.17	-1.25
Skewness	0.27	-0.02	0.59	-0.30	0.11	0.38	0.85	0.75	1.71	1.47	0.36
Range	6.63	3.47	17.00	11.00	16.65	25.00	84.60	7.00	3.00	8.00	1.21
Minimum	24.39	11.40	8.00	12.00	59.55	1.00	0.00	0.00	0.00	0.00	1.27
Maximum	31.02	14.87	25.00	23.00	76.19	26.00	84.60	7.00	3.00	8.00	2.48

Table 6 Descriptive statistics of studied parameters for Karnal bunt of wheat for 15 February-15 March month

Parameter	Tmax	Tmin	OTmax	OTmin	RHm	RHm>70	RF	RD	RD CNT	RD C10	HTR
Mean	23.98	10.98	10.58	16.25	71.76	16.50	33.53	3.17	0.67	2.42	2.27
Standard Error	0.54	0.41	1.75	1.43	1.55	2.27	8.88	0.81	0.28	0.73	0.14
Median	23.44	11.35	10.00	17.00	70.36	15.00	34.90	2.50	0.00	2.00	2.17
Standard Deviation	1.88	1.41	6.07	4.94	5.35	7.87	30.75	2.82	0.98	2.54	0.50
Sample Variance	3.52	1.99	36.81	24.39	28.65	61.91	945.51	7.97	0.97	6.45	0.25
Kurtosis	-0.87	-0.47	-1.31	0.14	-1.19	-1.02	-0.83	-1.14	1.70	-1.49	-1.93
Skewness	0.41	-0.47	0.01	-0.74	0.33	0.22	0.51	0.44	1.50	0.48	0.19
Range	6.13	4.75	18.00	17.00	15.78	25.00	88.20	8.00	3.00	6.00	1.26
Minimum	21.15	8.23	2.00	6.00	63.76	4.00	0.00	0.00	0.00	0.00	1.68
Maximum	27.28	12.98	20.00	23.00	79.53	29.00	88.20	8.00	3.00	6.00	2.95

During March, the mean minimum temperature was also higher (13.27°C) than that in February (9.07°C) month and 15 February-15 March period (10.98°C). The mean number of days when optimum minimum temperature prevailed were higher (17.92) during March as compared to15 February-15 March period (16.25) and February (10.00). Lower mean relative humidity was observed during the March (67.31%) month followed by 15 February-15 March period (71.76%) and February (73.80%). The mean number of days when mean relative humidity was higher (19.42) during February as compared to15 February-15 March period (16.50) and March (12.17). Lesser mean number of rainy days were observed during March (2.58) month as compared to 15 February-15 March period (3.17) and February (2.75). The mean number of weeks with continuous rainy days were higher for 15 February-15 March (0.67) period and equal for February and March month i.e. 0.58. The mean number of days when at least 10 mm rainfall was recorded were higher for 15 February-15 March (2.42) period as compared to February (2.33) and March (2.00) month. Lower mean HTR was observed for March (1.76) as compared to February (2.61) and 15 February-15 March period (2.27).

Development of disease prediction model

The results here are depicted here in both visual (Fig. 7) and numerical fashions (Tables 7, 8 and 9). The results demonstrate the adequacy of various methods of machine learning for prediction of Karnal bunt for different time period taken under study. The most intriguing finding is that for each period different models for disease prediction were perceived. The results accomplished surpass the earlier work in this area. The Taylor diagram (Fig. 7) provides readers with a comprehensive understanding of the degree of similarity between patterns in terms of correlation, root-mean-square difference, and variance ratio. While this diagram has a general application, it proves to be particularly valuable in assessing complex models.

As shown in Fig. 7, one can clearly see the observed or reference field, usually representing observed state. Another field is denoted as a test field usually representing model-simulated field. The purpose is to develop a theoretical framework of how closely the test field bear a resemblance to the reference field. The radial distances from the origin to the points represent pattern standard deviations. Correlation coefficient between two fields is illustrated by the azimuthal positions. The dashed lines represent RMSE values. For each period, cross-location multiple regression models {artificial neural networks (ANN), efficient neural network (ENET), k-nearest neighbour (kNN), least absolute shrinkage and selection operator (LASSO), Bayesian least absolute shrinkage and selection operator (BLASSO), support vector regression (SVR), ridge regression (RIDGE), Bayesian ridge (BRIDGE), multiple linear regression (MLR), principal component regression (PCR), and random forest (RF) approaches}were validated against each other. The models such as efficient neural network (ENET), k-nearest neighbour (kNN), Bayesian least absolute shrinkage and selection operator (BLASSO), support vector regression (SVR), Bayesian ridge (BRIDGE), principal component regression (PCR), and random forest (RF) for February

Fig. 7 Taylor diagram

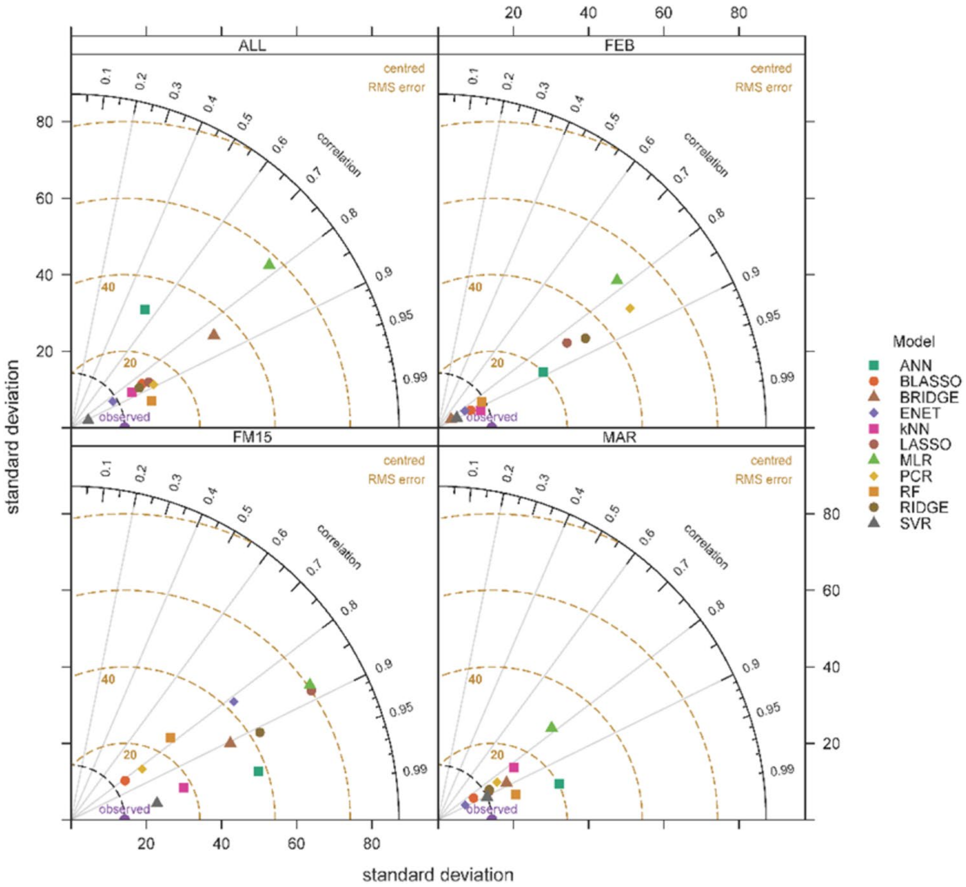

Table 7 Root mean square error (RMSE), Coefficient of determination (R^2) and Correlation coefficient (r) values of different machine learning models

Model	Root mean square error (RMSE)				Coefficient of determination (R^2)				Correlation coefficient (r)			
	FEB	MAR	FM15	ALL	FEB	MAR	FM15	ALL	FEB	MAR	FM15	ALL
ENET	14.48	15.10	46.92	15.91	0.73	0.78	0.66	0.73	0.85	0.88	0.81	0.85
MLR	51.83	30.91	60.55	58.23	0.60	0.61	0.76	0.61	0.78	0.78	0.87	0.78
RIDGE	35.32	16.86	43.78	18.04	0.74	0.75	0.83	0.75	0.86	0.86	0.91	0.87
BRIDGE	14.51	19.98	37.16	37.78	0.71	0.78	0.82	0.71	0.84	0.88	0.90	0.84
BLASSO	12.35	15.87	18.78	19.50	0.78	0.73	0.66	0.73	0.88	0.86	0.81	0.85
LASSO	31.73	17.90	59.91	21.91	0.70	0.77	0.78	0.75	0.84	0.88	0.88	0.87
ANN	29.02	29.42	34.26	38.06	0.79	0.92	0.94	0.29	0.89	0.96	0.97	0.54
SVR	10.99	8.97	18.91	12.15	0.79	0.83	0.97	0.85	0.89	0.91	0.98	0.92
RF	7.70	16.16	30.62	16.13	0.75	0.91	0.60	0.90	0.86	0.95	0.78	0.95
kNN	11.18	17.10	27.17	17.59	0.86	0.69	0.93	0.75	0.93	0.83	0.96	0.87
PCR	47.95	19.04	21.04	19.62	0.73	0.72	0.67	0.79	0.85	0.85	0.82	0.89

month perform relatively well because they lie relatively close to the reference point. Unlike others, the models such as artificial neural networks (ANN), k-nearest neighbour (kNN) and multiple linear regression (MLR) grossly underestimated the results for March month. Bayesian least absolute shrinkage and selection operator (BLASSO), support vector regression (SVR) and principal component regression (PCR) were reported as good models for 15 February-15 March period. All the models except efficient neural network (ENET) and k-nearest neighbour (kNN) grossly underestimated the results for overall period.

Validation of developed models

For February month, lower than mean RMSE, RRSE and MAE were observed for the models ENeT (14.48% ,117.11% and 12.89 %), BRIDGE (14.51%, 117.32 % and 12.33 %), BLASSO (12.35%, 99.86 % and 10.64 %), SVR (10.99%, 88.89 % and 9.76 %), RF (7.70%, 62.24 % and 6.20 %) and kNN (11.18%, 90.36 % and 10.10 %), respectively (Tables 7 and 8). The lowest RMSE, RRSE and MAE values were recorded for random forest (RF) model. The modified index of agreement (d) and modified Nash–Sutcliffe efficiency (E) values went maximum up to 0.73 and 0.44, respectively for RF model (Table 9). This indicates that d and E are not sensitive to systematic over or under prediction unlike other models. The coefficient of determination (R^2) and correlation coefficient (r) was highest for kNN model i.e., 0.86 and 0.93, respectively. For March month, the minimum RMSE, RRSE and MAE was observed for support vector regression (SVR) i.e., 8.97%, 72.51 % and 7.25%, respectively. But the coefficient of determination (R^2) and correlation coefficient (r) was highest for RF model i.e., 0.91 and 0.95, respectively and for SVR, R^2 and r were 0.83 and 0.91, respectively. The d and E values went maximum up to 0.71 and 0.34, respectively for SVR model that makes this criterion not much sensitive to quantification of systematic over or under prediction errors whereas the d and E values for RF model were quite less 0.57 and -0.28, respectively as compared to SVR model. SVR with RMSE value as 18.91%, RRSE as 152.91% and MAE value as 16.97% and BLASSO model with RMSE value as 18.78%, RRSE as 151.88% and MAE value as 16.53% perform relatively well as compared to other models for period 15 February to 15 March. The d values for BLASSO and SVR were 0.49 and 0.5, respectively and E values for BLASSO and SVR went maximum up to -0.50 and -0.54, respectively. But the coefficient of determination (R^2) and correlation coefficient (r) was highest for SVR model i.e., 0.97 and 0.98, respectively followed by ANN (R^2 = 0.94 and r=0.97). For overall period, little glitches

Table 8 Root relative square error (RRSE) and Mean absolute error (MAE) values of different machine learning models

Model	Root relative square error (RRSE) (%)				Mean absolute error (MAE) (%)			
	FEB	MAR	FM15	ALL	FEB	MAR	FM15	ALL
ENET	117.11	122.13	379.35	128.65	12.89	13.40	45.14	14.52
MLR	419.06	249.94	489.64	470.82	50.65	25.15	58.86	51.71
RIDGE	285.59	136.33	353.99	145.90	34.49	15.41	43.67	15.25
BRIDGE	117.32	161.57	300.51	305.46	12.33	17.81	36.84	36.48
BLASSO	99.86	128.34	151.88	157.65	10.64	14.48	16.53	16.30
LASSO	256.59	144.75	484.44	177.15	30.89	16.71	58.46	18.53
ANN	234.64	237.85	277.05	307.77	26.20	25.63	30.62	29.63
SVR	88.89	72.51	152.91	98.22	9.76	7.25	16.97	10.59
RF	62.24	130.65	247.62	130.46	6.20	14.10	22.10	13.73
kNN	90.36	138.29	219.68	142.21	10.10	12.09	22.64	15.54
PCR	387.74	153.93	170.13	158.63	47.06	16.98	17.15	17.38

Table 9 Modified Index of agreement (d) and Modified Nash–Sutcliffe efficiency (E) values of different machine learning models

Model	Modified Index of agreement (d)				Modified Nash–Sutcliffe efficiency (E)			
	FEB	MAR	FM15	ALL	FEB	MAR	FM15	ALL
ENET	0.49	0.45	0.30	0.48	-0.17	-0.22	-3.11	-0.32
MLR	0.28	0.44	0.25	0.28	-3.61	-1.29	-4.36	-3.70
RIDGE	0.36	0.49	0.31	0.55	-2.14	-0.40	-2.97	-0.39
BRIDGE	0.44	0.49	0.35	0.35	-0.12	-0.62	-2.35	-2.32
BLASSO	0.54	0.45	0.49	0.54	0.03	-0.32	-0.50	-0.48
LASSO	0.39	0.46	0.25	0.51	-1.81	-0.52	-4.32	-0.69
ANN	0.43	0.44	0.39	0.40	-1.38	-1.33	-1.79	-1.70
SVR	0.46	0.71	0.51	0.46	0.11	0.34	-0.54	0.04
RF	0.73	0.57	0.47	0.59	0.44	-0.28	-1.01	-0.25
kNN	0.58	0.62	0.47	0.53	0.08	-0.10	-1.06	-0.41
PCR	0.30	0.47	0.54	0.53	-3.28	-0.55	-0.56	-0.58

were observed. Lower RMSE, RRSE and MAE values were observed for SVR (12.15%, 98.22% and 10.59%), ENET (15.91%, 128.65% and 14.52%) and RF (16.16%, 130.46% and 13.71%) models. The corresponding d and E values were SVR (0.46 and 0.04), ENET (0.48 and -0.32) and RF (0.59 and -0.25). But the value of index of agreement (d) was higher for RF (0.59), RIDGE (0.55) and BLASSO (0.54) and Nash–Sutcliffe efficiency (E) values went maximum up to 0.04, -0.25 and -0.32 for SVR, RF and ENET models. The coefficient of determination (R^2) and correlation coefficient (r) was highest for RF model i.e., 0.90 and 0.95, respectively followed by SVR (R^2 = 0.85 and r = 0.92).

Table10 Tuning parameters of the machine learning models

Model	Tuning Parameter	FEB	MAR	FM15	ALL
ENET	Fraction	0.045	0.455	0.528	0.0455
	Lambda	0.002	0.002	1.986	0.003
MLR	AIC	104.56	100.63	77.97	928.27
RIDGE	Alpha	0	0	0	0
	Lambda	11.497	93.260	1.873	187.381
BRIDGE	NA	NA	NA	NA	NA
BLASSO	NA	NA	NA	NA	NA
LASSO	Alpha	1	1	1	1
	Lambda	3.274	11.497	0.024	10
ANN	Layer1	2	2	3	3
SVR	Sigma	0.1	0.1	0.1	0.1
	C	0.25	0.5	1	1
RF	mtry	3	3	3	3
kNN	k	11	2	3	6
PCR	N Component	3	3	1	2

Tuning parameters of the machine learning models

Tuning parameters are used in statistical modeling, particularly in shrinkage methods like RIDGE regression, LASSO regression, or Elastic Net (Table 10). They control the amount of shrinkage applied to model coefficients or data values. Shrinkage helps create simpler, more interpretable models and avoids overfitting when dealing with high-dimensional data or many predictors. The central point, often the mean, represents a prior belief about the data distribution. Shrinkage makes models more stable, robust, and better at generalizing to new data. It is valuable for limited data and problems with numerous predictors.

Conclusion

After rigorous investigation, key findings were emerged regarding the adequacy of various methods of machine learning for prediction of Karnal bunt for different time period taken under study. The most intriguing finding is that for each period, different models have performed well for disease prediction. The random forest regression (RF) for February month, support vector regression (SVR) for March month, SVR and BLASSO for 15 February to 15 March period and random forest for overall period surpassed the performance than other models. The suitability of these methods can be assessed for real time data and can be used for forewarning of Karnal bunt in Punjab.

Declarations

Conflict of interest The authors involved in this research declare no conflict of interest. They affirm that there are no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

Anbananthen KSM, Subbiah S, Chelliah D, Sivakumar P, Somasundaram V, Velshankar KH, Khan MKAA (2021) An intelligent decision support system for crop yield prediction using hybrid machine learning algorithms. F1000 Research 10:1–18. https://doi.org/10.12688/f1000research.73009.1

Anonymous (2019) Government of India. Kisan Knowledge Management System. https://dackkms.gov.in/account/login.aspx. Accessed 20 Sept 2023

Anonymous (2022). https://www.tutorialspoint.com/scikit_learn/scikit_learn_bayesian_ridge_regression.htm. Accessed 18 Sept 2023

Bhanumathi B, Vineeth M, Rohit N (2019) Crop Yield Prediction and Efficient use of Fertilizers. IEEE International conference on communication and signal processing (ICCSP), pp 769–773

Budhlakoti N, Rai A, Mishra DC (2020) Effect of influential observation in genomic prediction using LASSO diagnostic. Indian J Agric Sci 90(6):1155–9

Francl L (2001) The Disease Triangle: A plant pathological paradigm revisited. Plant Health Instr. https://doi.org/10.1094/PHI-T-2001-0517-01

Fu L, Feng Y, Majeed Y, Zhang X, Zhang J, Karkee M, Zhang Q (2018) Kiwi fruit detection in field images using Faster R-CNN with ZF Net. IFAC Pap 51:45–50

Garret KA, Dendy SP, Frank EE, Rouse MN, Travers SE (2006) Climate change effects on plant disease: genomes to ecosystems. Ann Rev Phytopath 44:489–509

Gonzalez-sanchez A, Frausto-solis J, Ojeda-bustamante W (2014) Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res 12(2):313–328

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT press

Gruber M (1998) Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators. Boca Raton: CRC Press, pp 7–15

Harefa J, Alexander A, Pratiwi M (2017) Comparison classifier: support vector machine (SVM) and K-nearest neighbor (K-NN) in digital mammogram images. J Informatika dan Sistem Informasi 2(2):35–40

Jhorar OP, Malvi HS, Sharma I, Mahi GS, Mathauda SS, Singh G (1992) A biometeorological model for forecasting Karnal bunt disease of wheat. Plant Dis Res 7:204–9

Jin W, Li ZJ, Wei LS, Zhen H (2000) "The improvements of BP neural network learning algorithm", In WCC 2000-ICSP 2000. 2000 5th international conference on signal processing proceedings. 16th world computer congress. IEEE 3:1647–1649

Kaur S, Singh K (2000) Effect of seasonal variations in temperature and relative humidity on the development of Karnal bunt of wheat. J Res Punjab Agric Univ 37:71–7

Kaur G, Kaur S, Dhaliwal LK (2006) Response of commercial old and new varieties from the region of wheat to yellow rust and Karnal bunt. J Res Punjab Agric Univ 43:316–22

Kaur J, Bala R, Kaur H, Pannu PPS, Kumar A, Bhardwaj SC (2018) Current status of wheat diseases in Punjab. Agric Res J 55:113–6

Kennedy P (2003) A guide to econometrics (Fifth ed). Cambridge: The MIT Press, pp 205–206

Mavi SS, Jhorar OP, Sharma I, Mahi GS, Mathauda SS, Aujla SS (1992) Forecasting Karnal bunt disease of wheat-agronomical method. Cereal Res Commun 20:744–67

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — A discussion of principles. J Hydrol 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Newbery F, Qi A, Fitt BD (2016) Modelling impacts of climate change on arable crop diseases: Progress, challenges and applications. Curr Opin Plant Biol 32:101–109

Park T, Casella G (2008) The bayesian lasso. J Amer Stat Assoc 103(482):681–686

Paszke A, Chaurasia A, Kim S, Culurciello E (2016) ENet: a deep neural network architecture for real-time semantic segmentation. 1–10. http://arxiv.org/abs/1606.02147. Accessed 12 Sept 2023

Prakash JS, Vignesh KA, Ashok C, Adithyan R (2012) Multi class support vector machines classifier for machine vision application. In: 2012 International conference on machine vision and image processing (MVIP). IEEE, pp 197–199

Quinlan JR (1992) Learning with continuous classes. Proc. AI'92, 5th Aust. Joint Conf. on Artificial Intelligence (Adams & Sterling, eds.). World Scientific, Singapore, pp 343–348

Sandhu SK, Attri A, Bala R (2022) Effect of meteorological parameters on Karnal bunt incidence in wheat under different agroclimatic zones of Punjab. J Agrometeorol 24:66–71

Sharma I, Bains NS, Sharma RC (2012) Resistance in wheat to Karnal bunt. In: Sharma I (ed) Disease Resistance in wheat. CAB International, UK, pp 190–220

Singh D, Singh R, Rao V, Karwasra SS, Beniwal MS (1996) Relation between weather parameters and Karnal bunt (*Neovossia indica*) in wheat (*Triticum aestivum*). Ind J Agric Sci 66:522–5

Smilanick JL, Prescott JM, Hoffmann JA, Secrest LR, Wiese K (1989) Environmental effects on survival and growth of secondary sporidia and teliospores of *Tilletia indica*. Crop Prot 8:86–90

Smiley RW, Patterson LM (1996) Pathogenic fungi associated with Fusarium foot rot of winter wheat in the semiarid Pacific Northwest. Plant Dis 80:944–949

Srinivasan G (1980) Role of meteorological factors in the epidemiology of Karnal bunt and rust disease of wheat. M.Sc. Thesis. Dept of Agronomy and Agrometeorology, PAU, Ludhiana

Stevens R (1960) An advanced treatise. Plant Pathol. 3:357–429

Tatem AJ, Rogers DJ, Hay SI (2006) Global transport networks and infectious disease spread. Adv Parasitol 62:293–43

Techopedia (2020) Artificial Neural Network (ANN). https://www.techopedia.com/definition/5967/artificial-neural-network-ann. Accessed 20 Jan 2020

Waglea SA, Harikrishnan R (2022) Prediction of tomato plant disease with meteorological condition and artificial intelligence. ECS Transactions 107(1):20377–84

Wang Q, Zhang T, Cui J, Wang X, Zhou H, Han J, Gislum R (2011) Path and ridge regression analysis of seed yield and seed yield components of russian wildrye (*Psathyrostachys juncea nevski*) under field conditions. PLoS ONE 6(4):1–10. https://doi.org/10.1371/journal.pone.0018245

Willmot CJ (1984) On the evaluation of model performance in physical geography. In: Spatial statistics and models, edited by: Gaile, G. L. and Willmot, C. J., D. Reidel. Dordrecht, pp 443–460

Winston PH (1992) Artificial intelligence. Addison-Wesley Longman Publishing Co., Inc