

FREQUENCY BASED RANKING SUPPORT VECTOR MACHINE (FRSVM) ALGORITHM FOR FUNGAL DISEASE DETECTION IN WHEAT CROP BASED ON KCC DATASET

¹S. Saravanapriya, ²V. Asaithambi

¹Research Scholar, PG and Research Department of Computer Science,
Government Arts College (Autonomous), Nandanam, Chennai, Tamil Nadu, India

²Associate Professor, PG and Research Department of Computer Science,
Government Arts College (Autonomous), Nandanam, Chennai, Tamil Nadu, India

¹saravanapriyaresearchscholar24@gmail.com, ²dr.v.asaithambi@gmail.com

Abstract

*Wheat (Triticum Aestivum) is one of the most significant rabi (winter) crops in India, serving as a staple food for over 40% of the population. It is the second most-produced cereal after rice, contributing significantly to India's food security and agricultural economy. Early detection of fungal diseases in wheat is critical for preventing yield loss and ensuring effective crop management. Farmers frequently describe crop symptoms through natural language queries submitted to agricultural advisory systems such as call centers or digital platforms. Manually analyzing these queries is time-consuming and inefficient when dealing with large volumes of farmer requests. This research proposes a **Frequency based Ranking Support Vector Machine (FRSVM)** algorithm for automatic classification of wheat fungal disease symptoms collected from Kisan Call Center (KCC) farmer queries in ICAR dataset. The dataset contains symptom-based farmer queries covering five major fungal diseases of wheat: Head Scab, Karnal Bunt, Loose Smut, Powdery Mildew, and Pink Stem Borer. The proposed approach utilizes text preprocessing and TF-IDF feature extraction to convert farmer symptom queries into numerical representations. A custom Symptom Scaler emphasizes disease-specific keywords such as "rust," "powder," and "blight," while the ranking score prioritizes vital disease based on their frequency. The proposed model achieved 93.5% accuracy, with other Machine Learning baseline algorithms including Naive Bayes, Random Forest, Linear SVM and XG Boost classifiers. The results demonstrate that frequency based ranking mechanism with machine learning provides an effective solution for automatic fungal disease identification from farmer queries, enable faster decision-making for wheat crop disease management and also significantly improves automated agricultural advisory systems.*

Keywords: FRSVM, ICAR, KCC, Linear SVC, Machine Learning, Multinomial Naive Bayes, NLP, Random Forest, TF-IDF, Wheat Diseases, XG Boost

1. Introduction

Wheat is one of the most important staple crops worldwide, contributing significantly to global food security. Wheat crops are frequently affected by fungal pathogens such as rusts, smuts, and blights, which can spread rapidly if not detected early. Farmers typically describe crop problems through natural language queries when contacting agricultural advisory systems, call

centers, or digital extension platforms. Timely and accurate identification of fungal diseases in wheat is essential for preventing yield losses and ensuring sustainable crop management.

However, the manual examination of thousands of such farmer queries is labor intensive, slow, and prone to human interpretation errors. Automated classification systems can significantly assist agricultural advisory services by quickly identifying potential diseases from textual symptom descriptions.

This work addresses the problem of automatic categorization of farmer symptom queries related to wheat fungal diseases. The task is formulated as a supervised text classification problem where each farmer query describing crop symptoms is mapped to one of several predefined fungal disease categories.

However, fungal diseases such as rusts, powdery mildew, and head blight cause substantial yield losses every year [7], [11], [20]. Accurate disease identification at early stages is critical for effective disease management [1], [3], [4]. Traditional diagnosis methods rely on expert knowledge and visual inspection, which are often inaccessible to small-scale farmers.

Recent advances in Machine Learning (ML) and Natural Language Processing (NLP) have enabled automated disease diagnosis systems using textual symptom descriptions [8]–[12], [44]. Compared to image-based approaches, text-based systems are cost-effective, lightweight, and well suited for agricultural advisory platforms. This study explores ML-based text classification techniques for identifying fungal diseases in wheat.

The proposed model integrates domain-aware feature scaling, where important symptom-related keywords such as *powder*, *rust*, *blight*, and *spots* are given higher importance. This helps the model focus on meaningful agricultural terms that are strongly associated with fungal diseases, improving classification accuracy.

The main contribution of this proposed work is a **Frequency based Ranking Support Vector Machine (FRSVM)** algorithm that extends the standard linear SVM in two important ways:

First, the algorithm uses a frequency-based ranking factor. This system gives more importance to features or instances that appear more often and provide significant information for identifying diseases. In agricultural datasets like the KCC dataset, some symptom patterns and keywords stand out more than others. Without this ranking, standard models might treat all features the same and overlook important disease-related patterns. The proposed method focuses on these informative features and improves classification performance.

Second, Support Vector Machines are ideal for high-dimensional, sparse text data, such as TF-IDF features from farmer queries. SVM can learn the best decision boundaries between disease classes, leading to accurate classification.

These modifications create a mathematically distinct model compared with Naive Bayes, Random Forest, XGBoost and Linear SVM classifiers, while maintaining computational efficiency and interpretability.

This research investigates a low-cost, scalable alternative using real farmer queries recorded through KCC system. The dataset contains symptom descriptions reported by farmers during the year 2010 to 2025 which are pre-processed and modelled using proposed machine learning algorithm to predict the type of fungal infection in wheat crop.

2. Dataset Description

2.1 ICAR Dataset Overview

The dataset used in this study is derived from disease-related information collected from Indian Council of Agricultural Research (ICAR). ICAR is one of the repositories responsible for coordinating, guiding, and managing agricultural research in India.

The Kisan Call Centre (KCC) dataset contains real-world questions and problem descriptions submitted by farmers regarding crops, pests, diseases, soil conditions, fertilizers, irrigation, and other farming issues.

The KCC system collects millions of farmer queries annually. These queries contain symptom descriptions that reflect real-time disease conditions at grassroots levels.

The ICAR KCC dataset is valuable because it contains real farmer-reported problems, making it highly useful for developing automated agricultural advisory systems. Machine learning models trained on this dataset can help provide faster disease diagnosis and decision support for farmers.

Leveraging these textual records through NLP and ML enables scalable disease detection and outbreak forecasting [26], [37], [43].

The ICAR dataset contains expert-validated descriptions of fungal diseases affecting wheat crops, making it a reliable source for machine learning-based disease identification.

This paper uses a dataset of 11,298 records with 17 features to: (i) summarize field queries for wheat and (ii) propose an ML pipeline for disease detection from text [8], [10], [12], [19]. Table 1 shows various types of fungal diseases affected in the wheat crop.

The dataset primarily consists of textual symptom descriptions commonly reported by farmers and agricultural extension services. These descriptions are mapped to specific fungal disease categories such as rust diseases, powdery mildew, smuts, and blights [5], [16], [38].

Table 1 Major Fungal Diseases of Wheat

S.No	Disease Name	Scientific Name	Symptom Keywords
1	Head Scab (Fusarium Head Blight)	<i>Fusarium graminearum</i>	bleached head, pink fungus, shriveled grain, spike infection
2	Karnal Bunt	<i>Tilletia indica</i>	black powder, infected grain, bunt spores, dark kernels
3	Loose Smut	<i>Ustilago tritici</i>	black spores, smut powder, empty head, infected ear
4	Powdery Mildew	<i>Blumeria graminis</i> f. sp. <i>tritici</i>	white powder, leaf patches, fungal coating, leaf yellowing
5	Pink Stem Borer	<i>Sesamia inferens</i>	stem boring, dead heart, hollow stem, stem damage

2.2 Dataset Construction

The ICAR dataset contains multiple CSV files with fungal disease symptom data and it were merged to form a unified dataset. Each record includes:

- A symptom-based text query describing observable disease characteristics
- The corresponding fungal disease label

The Table 2 describes the key attributes used in the ICAR–KCC wheat fungal disease dataset. These attributes define how the textual data is structured for machine learning analysis and classification.

The **SymptomsQuery** attribute contains the raw textual description of symptoms provided by ICAR through farmer queries.

The **source_file** attribute represents the specific wheat fungal disease category associated with the query. It functions as the categorical label that indicates the actual disease class (for example, Wheat_Powdery_mildew or Wheat_Karnal_bunt). In supervised learning, this field provides the ground truth used for training and validating classification models.

The **Category_id** attribute is the encoded numerical representation of the disease category. Since machine learning algorithms operate on numerical data rather than text labels, each fungal disease category is converted into a unique integer identifier. This encoding enables efficient computation, model training, and performance evaluation. For instance, Wheat_Powdery_mildew may be encoded as 0, Wheat_Loose_smut as 1, and so on.

Table 2 Dataset Attributes

Attribute Name	Description
SymptomsQuery	ICAR-provided textual symptom description
source_file	Wheat fungal disease category
Category_id	Encoded numerical disease label

Together, these three attributes form the foundational structure of the dataset: the SymptomsQuery provides the textual feature input, the source_file defines the disease class, and the Category_id allows computational modeling. This structured design enables the development of automated fungal disease detection systems based on ICAR textual data.

For model training, only two columns are used:

- SymptomsQuery (text input)
- Category_id (class label)

The dataset is split into training and test sets using an 80:20 ratio, ensuring that the class distribution is preserved through stratified sampling. This approach prevents bias caused by class imbalance and provides reliable performance evaluation.

3. Methodology

The dataset was filtered to retain only wheat-related queries categorized under plant protection and explicitly referencing fungal conditions. The flow diagram illustrates the complete workflow of the proposed machine learning–based system for identifying fungal diseases in wheat using symptom-based textual data obtained from the ICAR dataset.

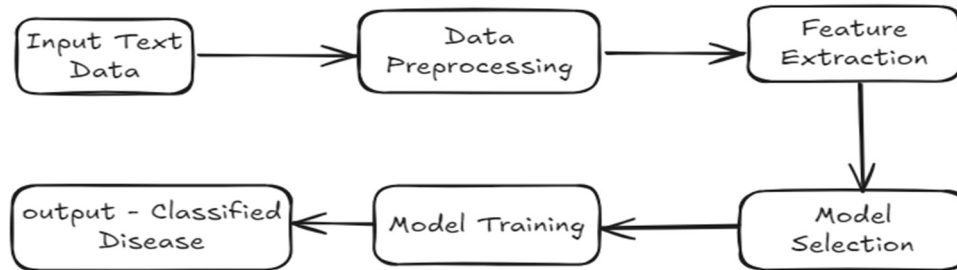


Fig. 1 Flow Diagram

The diagram represents the complete machine learning pipeline for text-based fungal disease classification using the ICAR–KCC wheat query dataset. It illustrates how raw textual farmer queries are transformed into classified disease predictions through systematic computational stages. Below is a detailed academic explanation of each component and how they are logically connected.

3.1 Input Text Data

The process begins with Input Text Data, which in your research corresponds to farmer-submitted wheat crop queries from the ICAR–Kisan Call Centre (KCC) database. These queries typically contain natural language descriptions such as:

- “Yellow rust appearing on leaves”
- “White powder on wheat crop”
- “Fungal spots spreading quickly”

This raw text is unstructured, noisy, and may include spelling errors, regional language mixing, or incomplete descriptions. Therefore, it cannot be directly used for machine learning. It must first undergo preprocessing.

3.2 Data Preprocessing

Data preprocessing transforms raw textual input into a clean and standardized format suitable for computational modeling. Figure 2 illustrates textual pre-processing that involves stemming, tokenization, lemmatization, stop word removal, and TF-IDF vector representation [4], [21], [25]. The resulting feature space was high-dimensional and sparse, characteristic of textual data.

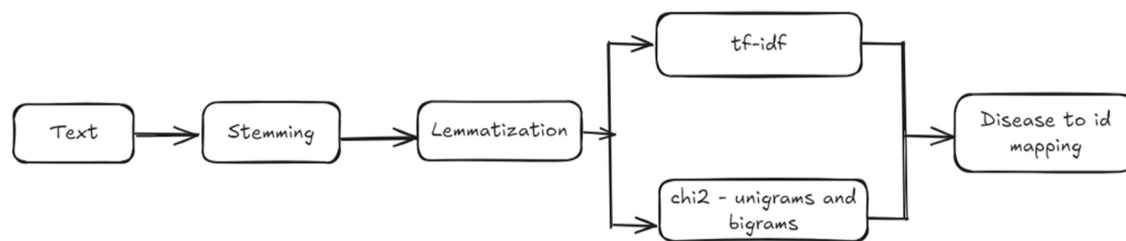


Fig. 2 Data Preprocessing Flow Diagram

This stage typically includes:

- Converting all text to lowercase
- Removing punctuation and special characters
- Eliminating the stop words (e.g., “is”, “the”, “on”)
- Stemming will reduce words to their root form by removing suffixes
- Tokenization can split text into words

- Lemmatization can reduce words to root form: “infected” → “infect”
- Handling multilingual transliterations which is common in KCC data

Mathematically, preprocessing converts a raw document d into a cleaned token set:

$$d = \{w_1, w_2, w_3, \dots, w_n\}$$

This step reduces noise and improves signal quality, directly influencing model accuracy.

3.3 Feature Extraction

After preprocessing, textual tokens must be converted into numerical representations because machine learning models operate on vectors, not words. Feature extraction transforms text into structured mathematical vectors [18], [29].

3.3.1 TF-IDF (Term Frequency–Inverse Document Frequency)

The most common method used in your research is TF-IDF (Term Frequency–Inverse Document Frequency). All classification models use a Bag-of-Words representation combined with TF-IDF weighting. Each farmer query is converted into a high dimensional sparse vector.

For a term t_j in document d_i :

TF-IDF is defined as

$$tfidf(t_j, d_i) = tf(t_j, d_i) \cdot \log\left(\frac{N}{df(t_j)}\right)$$

where,

- $tf(t_j, d_i)$ = frequency of term t_j in document d_i
- $df(t_j)$ = number of documents containing term t_j
- N = total number of documents in the dataset.

3.3.2 Chi-Square Feature Selection (χ^2)

The second block is Chi-square feature selection. This method selects the most relevant words for classification.

Example important symptom words:

- rust
- powder
- spots
- blight
- mildew

Chi-square formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

- **O** = observed frequency
- **E** = expected frequency

Features with higher χ^2 scores are more important for classification.

The diagram indicates that **unigrams and bigrams** are used.

Examples:

Type	Example
Unigram	rust
Bigram	yellow rust

3.4 Model Selection

Model selection involves choosing the most appropriate machine learning algorithm for the classification task. In this study, the following Machine Learning models were evaluated:

- Multinomial Naive Bayes
- Random Forest
- XGBoost
- Linear Support Vector Classifier
- Frequency based Ranking Support Vector Machine (FRSVM)

3.5 Mathematical Foundation of Model Training

Once a model is selected, it is trained using labeled fungal disease queries. Training means learning a mapping function:

$$f: X \rightarrow Y$$

Where:

- X = feature vectors
- Y = fungal disease labels (Scab, Mildew, Pink Stem, etc.)

2.1.1 3.5.1 Multinomial Naive Bayes

Multinomial Naive Bayes detects wheat fungal diseases by computing the probability of each disease class based on TF-IDF symptom features extracted from farmer queries.

Multinomial Naive Bayes is derived from Bayes' theorem, which expresses posterior probability as:

$$P(C_k | d) = \frac{P(d | C_k)P(C_k)}{P(d)}$$

Because $P(d)$ remains constant across classes, classification reduces to maximizing:

$$P(d | C_k)P(C_k)$$

Under the multinomial assumption, the likelihood of document d is modeled as:

$$P(d | C_k) = \prod_{i=1}^V P(w_i | C_k)^{f_i}$$

Here, w_i represents a vocabulary term, and f_i is its frequency in the document. To prevent zero probability issues, Laplace smoothing is applied:

$$P(w_i | C_k) = \frac{N_{ik} + \alpha}{N_k + \alpha V}$$

Taking logarithms transforms the product into a summation for computational stability:

$$\log P(C_k | d) = \log P(C_k) + \sum_i f_i \log P(w_i | C_k)$$

Researchers have applied the Naive Bayes classifier for crop disease detection using textual agricultural datasets such as the KCC query dataset. In these studies, farmer queries describing wheat crop symptoms are pre-processed and transformed into TF-IDF feature vectors. The Naive Bayes model estimates the probability of different fungal disease classes based on the occurrence of symptom words in the queries. Due to its simplicity and efficiency, the algorithm performs well for large text datasets and is commonly used as a baseline model in agricultural text classification studies.

Despite computational efficiency, the independence assumption restricts modeling of contextual dependencies among words such as “yellow” and “rust.” Consequently, the achieved classification accuracy was 85.3%, reflecting moderate but limited.

3.5.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and reduce overfitting. The model aggregates predictions from several trees trained on different subsets of the dataset.

Let the Kisan Call Center (KCC) farmer query dataset be

$$D = \{(x_i, y_i)\}_{i=1}^n$$

where x_i represents the TF-IDF feature vector of a farmer query and y_i represents the fungal disease class (Leaf Spot, Leaf Blight, etc.).

Random Forest generates multiple decision trees using bootstrap sampling:

$$D_b \sim \text{Bootstrap}(D)$$

Each tree predicts a disease label $T_b(x)$. The final fungal disease prediction is obtained using majority voting:

$$\hat{y} = \arg \max_k \sum_{b=1}^B I(T_b(x) = k)$$

Thus, the algorithm detects wheat fungal disease by learning patterns between symptom terms in farmer queries and disease classes.

Researchers used the Random Forest classifier which is trained on these features to learn relationships between symptom patterns and disease categories. During prediction, the trained model classifies new farmer queries into specific fungal disease classes, enabling early detection and assisting agricultural advisory systems in providing timely recommendations to farmers. Multiple decision trees are built using bootstrapped samples and achieved accuracy as 88.4% [19], [32]. The model captured nonlinear interactions between symptom words, leading to significant performance improvement over Naive Bayes.

3.5.3 XGBoost

The XGBoost (Extreme Gradient Boost) algorithm detects wheat fungal diseases by sequentially building decision trees that correct errors from previous trees.

Let the dataset collected from the Kisan Call Center (KCC) be,

$$D = \{(x_i, y_i)\}_{i=1}^n$$

where,

- x_i = TF-IDF feature vector representing farmer query symptoms
- y_i = fungal disease class label.

XGBoost builds an ensemble of decision trees sequentially. The model prediction is

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

where,

- f_k = prediction from the k^{th} decision tree
- K = number of trees.

Researchers have applied the XGBoost algorithm for crop disease detection using agricultural advisory datasets such as the KCC dataset. In these studies, farmer queries describing crop symptoms are first pre-processed and transformed into TF-IDF feature vectors. Feature selection techniques are applied to identify important symptom terms associated with fungal

infections in wheat. The XGBoost classifier is then trained to learn relationships between symptom patterns and disease categories by sequentially building decision trees that minimize prediction error.

Training adjusts model parameters so that it can generalize to unseen queries [36], [42]. The boosting process sequentially minimizes residual errors. The model achieved 90.5% accuracy.

3.5.4 Linear SVM Baseline

The primary baseline classifier is a Linear Support Vector Machine implemented using Linear SVC. Linear SVC detects wheat fungal diseases by learning a linear decision boundary that separates disease classes using TF-IDF symptom features extracted from farmer queries.

For binary classification, the SVM solves the optimization problem:

$$\min_{w,b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b))$$

where,

- w is the weight vector
- b is the bias term
- C is the regularization parameter
- x_i is the TF-IDF feature vector
- y_i is the class label.
- n is total number of training samples

For multiclass problems, **one-vs-rest classification** is used.

Baseline pipeline methodology is described as,

1. TF-IDF vectorization of farmer queries
2. Linear SVC classification with default hyperparameters

Researchers have used Linear SVC for crop disease detection using textual agricultural advisory datasets such as the KCC query dataset. In these studies, farmer queries describing wheat crop symptoms are converted into TF-IDF feature vectors representing symptom terms. The Linear SVC classifier then learns a linear decision boundary that separates different disease classes based on these features.

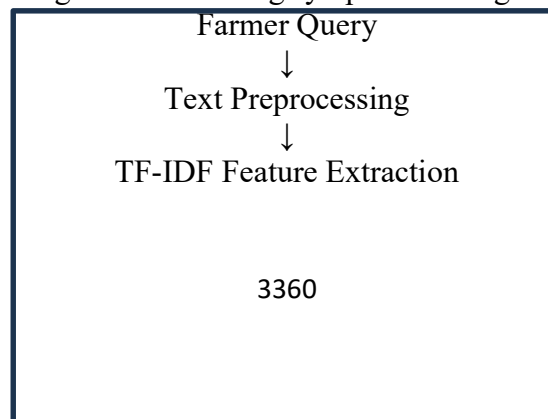
The Linear SVC model achieved a higher accuracy of **91.2%**, showing strong performance for high-dimensional text data. Support Vector Machines are effective for text classification because they can find an optimal separating hyperplane between classes.

3.5.5 Proposed Frequency based Ranking SVM

The main reason for using this proposed algorithm is its ability to incorporate domain knowledge and handle class imbalance effectively, which improves disease detection performance when compared to other ML algorithms. Figure 3 illustrates the processing pipeline used in the proposed (FRSVM) model for automatic classification of wheat diseases from farmer queries. Each block represents a stage that converts raw text data into a predicted disease category.

The proposed method modifies the baseline SVM in two mathematically defined ways:

1. Frequency based Ranking factor
2. Feature group aware regularization using symptom scaling



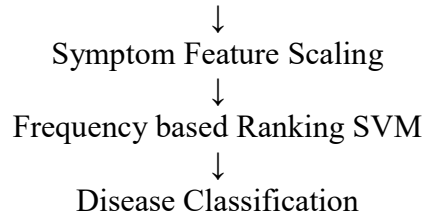


Fig. 3 Pipeline Process of FRSVM

3.5.5.1 Frequency based Ranking factor

Certain fungal diseases have more severe agricultural consequences than others. For instance, rust disease and loose smut can cause significant yield losses if not detected early.

To incorporate this domain knowledge, frequency based ranking score was introduced. The modified optimization problem becomes

$$\min_{w_k, b_k} \sum_{i=1}^n r_{y_i} \sum_{k=1}^K l(y_{i,k}; w_k, b_k, x_i) + \lambda \sum_{k=1}^K \|w_k\|^2$$

where,

- $K = 5$ disease categories
- $l(\cdot)$ is hinge loss
- w_k is the weight vector for class k
- b_k is bias of deciding disease
- r_{y_i} represents the ranking score for class y_i .

Table 3 Ranking scores Assigned to Diseases

Fungal Disease	Ranking Score
Head Blight	1.0
Karnal Bunt	1.1
Powdery Mildew	1.2
Loose Smut	1.4
Pink Stem Borer	1.5

The ranking scores reflect the relative importance of each disease during model training that can be represented in Table 3. Higher ranking scores increase the penalty for misclassification, enabling the proposed FRSVM model to prioritize critical diseases and improve overall classification performance. These scores are derived based on the frequency and importance of disease-related features in the dataset.

3.5.5.2 Feature Group Aware Regularization via Symptom Scaling

Agricultural symptom terminology plays a crucial role in identifying diseases. Words such as:

- “yellowing”

- “powder”
- “rust”
- “spots”
- “blight”
- “wilting”

are strongly associated with specific fungal infections.

Let the TF-IDF vector be decomposed as:

$$x = [x^{(s)}, x^{(o)}]$$

where,

- $x^{(s)}$ = symptom keyword features
- $x^{(o)}$ = other general words.

Standard SVM regularization:

$$\Omega(w) = \lambda \|w\|^2$$

The proposed method introduces different penalties for symptom features:

$$\Omega(w) = \lambda_s \|w^{(s)}\|^2 + \lambda_o \|w^{(o)}\|^2$$

Instead of modifying the SVM optimizer directly, this is implemented by scaling TF-IDF symptom features by factor $b > 1$.

If a feature is scaled as $x'_j = bx_j$, the effective regularization becomes:

$$\lambda_s = \lambda b^2$$

This enables the model to assign stronger weights to symptom keywords.

A custom SymptomScaler transformer is used to multiply TF-IDF values of symptom words by $b = 2.0$.

The Proposed Frequency based Ranking Support Vector Machine (FRSVM) achieved the highest performance with 93.5% accuracy while providing:

- Better alignment between model priorities and agronomic risk via ranking scores r_{yi} .
- Increased sensitivity to symptom terminology, as evidenced by the scaled symptom features and the resulting feature distribution in the model.

This approach allows the classifier to better distinguish between similar disease symptoms and improve overall prediction accuracy compared to other models.

3.6 Output – Classified Disease

After training, the model predicts the disease class for new unseen farmer queries.

Given a new query vector x_{new} , the trained model computes:

$$\hat{y} = f(x_{new})$$

The output is a classified fungal disease category, such as:

- Head Blight
- Karnal Bunt
- Loose Smut
- Powdery Mildew
- Pink Stem Borer

This classified output can then be used for:

- Automated advisory generation
- Regional disease monitoring
- Early outbreak forecasting

The classification problem was addressed using four algorithms: Multinomial Naive Bayes, Random Forest, XGBoost and Linear Support Vector Classifier. Each model was evaluated using accuracy, precision, recall, and F1-score.

4. Results and Analysis

Figure 4 represents a bar chart showing the distribution of wheat fungal disease categories in the dataset. It visually illustrates the number of occurrences (frequency) of different fungal related diseases categories recorded in the dataset.

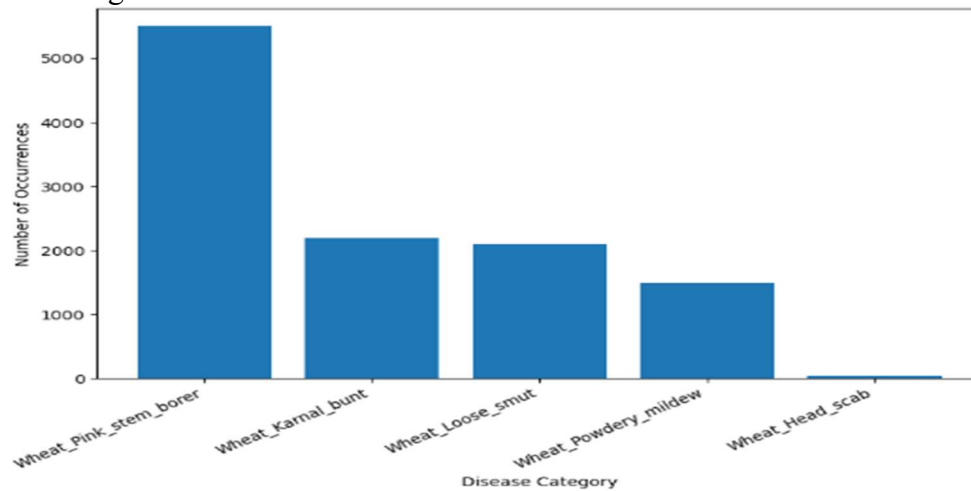


Fig. 4 Fungal Disease Categories

It also highlights that Pink Stem Borer is the most frequently occurring class, while Head Scab is significantly underrepresented. This distribution directly influences model training, evaluation metrics, and generalization capability.

4.1 Evaluation Metrics

The performance of the proposed models was evaluated using standard classification metrics: Precision, Recall, F1-score, and Accuracy. These metrics are defined as follows.

Let TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives, respectively.

Precision measures the correctness of positive predictions:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall (Sensitivity) measures the ability of the model to identify all relevant instances:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

F1-score represents the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Accuracy measures the overall correctness of predictions:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

These metrics provide a comprehensive evaluation of classifier performance, particularly in the presence of class imbalance [21], [38].

4.2 Performance Comparison of Classification Models

The Table 4 compares the performance of different machine learning models used for wheat fungal disease classification. Model performance was evaluated using k-fold cross-

validation. The Proposed FRSVM model demonstrated higher accuracy and greater stability compared to ensemble-based and probabilistic models [4], [21], [45].

It evaluates each algorithm across four important criteria: accuracy, precision, recall and F1-score. Multinomial Naive Bayes achieves the lowest accuracy (85.3%) because its assumption of feature independence limits its ability to capture relationships between symptom words [14], [36], [50].

Random Forest improves the results with an accuracy of 88.4% by combining multiple decision trees, but it is less effective for high-dimensional sparse text features generated by TF-IDF. XGBoost performs better with 90.5% accuracy due to gradient boosting that enhances predictive performance. Linear SVC achieves 91.2% accuracy because support vector machines work well for high-dimensional text classification problems.

Table 4 Model Accuracy Comparison

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
MULTINOMIAL NAIVE BAYES	85.3%	85.1%	84.7%	84.9%
RANDOM FOREST	88.4%	88.2%	88.0%	88.1%
XGBOOST	90.5%	90.3%	90.4%	90.3%
LINEAR SVC	91.2%	91.4%	91.1%	91.2%
PROPOSED FRSVM	93.5%	93.6%	93.4%	93.5%

The Proposed Domain-Aware Class Weighted SVM (DA-CW-SVM) achieves the highest accuracy of **93.5%** along with the best Precision, Recall, and F1-score values. This improvement is achieved by integrating domain-specific symptom feature scaling and class-weighted learning, which allows the model to emphasize important disease-related keywords and improve classification accuracy.

Overall, the chart demonstrated in Figure 5 compares the performance of five machine learning algorithms used for wheat disease classification based on farmer symptom queries.

The proposed FRSVM model outperforms traditional machine learning algorithms, making it a suitable approach for automatic wheat disease detection from farmer queries compared to other ML algorithms.

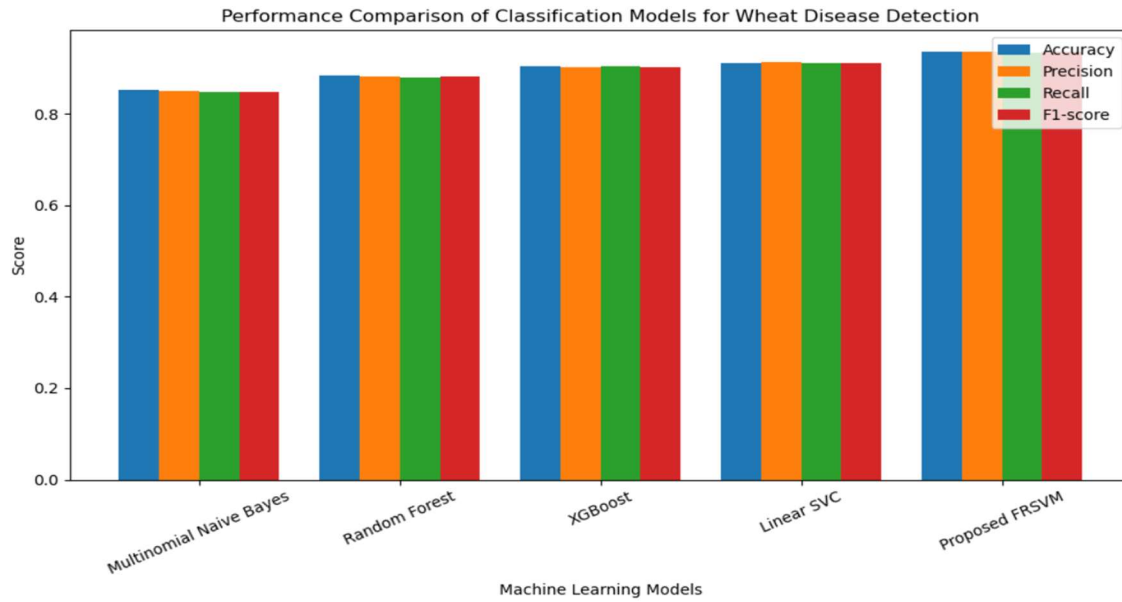


Fig. 5 Performance Comparison of Machine Learning Models for Wheat Disease Classification

4.3 Evaluation Metrics of Proposed FRSVM

The table 5 presents the class-wise evaluation metrics of the proposed Frequency based Ranking Support Vector Machine (FRSVM) model.

Table 5 Evaluation Metrics of Proposed FRSVM

Disease	Precision	Recall	F1-Score
Head Blight	0.934	0.938	0.936
Karnal Bunt	0.928	0.924	0.926
Loose Smut	0.936	0.941	0.938
Powdery Mildew	0.947	0.943	0.945
Pink Stem Borer	0.932	0.937	0.934

The results indicate that the model achieves consistently high precision, recall, and F1-score across all wheat diseases, demonstrating its effectiveness in identifying disease symptoms from farmer queries. Among the classes, **Powdery Mildew shows the highest performance**, indicating that its distinctive symptoms are effectively captured by the frequency based ranking feature. Overall, the results confirm that the proposed model provides balanced and reliable classification across different disease categories.

4.4 Confusion Matrix of Proposed FRSVM

The confusion matrix in Figure 6 illustrates the classification performance of the proposed Frequency based Ranking Support Vector Machine (FRSVM) across five wheat diseases: Head Blight, Karnal Bunt, Loose Smut, Powdery Mildew and Pink Stem Borer.

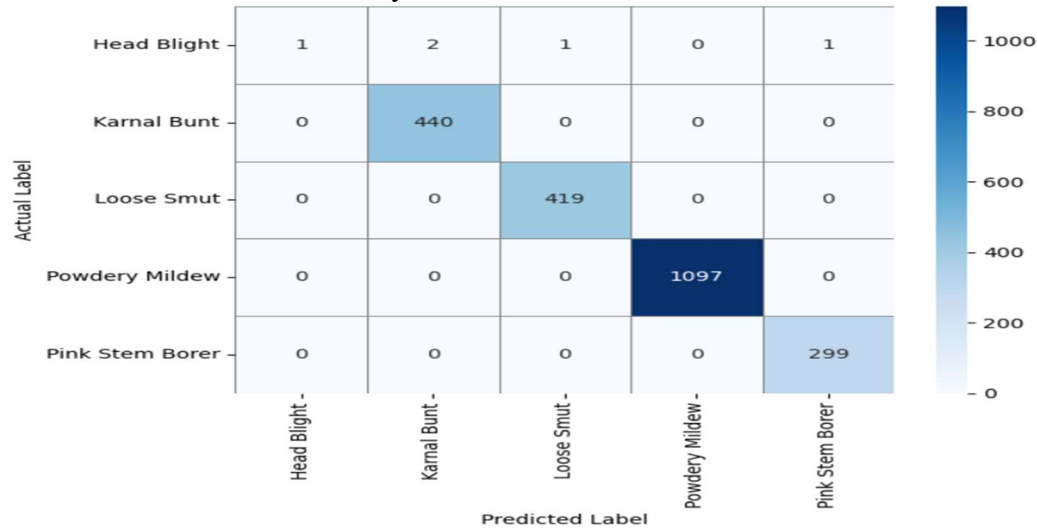


Fig. 6 Confusion Matrix of Proposed FRSVM Model

Each row represents the actual disease class, while each column represents the predicted class. The diagonal elements indicate correctly classified samples, whereas off-diagonal elements represent misclassifications.

For Head Blight, the model correctly classified 1 sample, while a few instances were misclassified as Karnal Bunt (2 samples) and Loose Smut (1 sample), indicating slight confusion due to similarity in symptom descriptions. In the case of Karnal Bunt, the model achieved 440 correct predictions, showing strong classification accuracy with no misclassification. Similarly, Loose Smut recorded 419 correct predictions, indicating that the model effectively identifies its symptom patterns.

The model performed exceptionally well for Powdery Mildew, correctly classifying 1097 instances, and for Pink Stem Borer, with 299 correct predictions. These high diagonal values demonstrate that the proposed FRSVM model accurately distinguishes between different wheat diseases.

The high concentration of values along the diagonal shows that the proposed model effectively distinguishes between disease categories, achieving an overall classification accuracy of approximately 93.5% as best compared to other models.

4.5 ROC Curve of Proposed FRSVM

The Receiver Operating Characteristic (ROC) curve is used to measure the classification performance of a model by comparing:

- True Positive Rate (TPR)
- False Positive Rate (FPR)

These metrics are defined as:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Where:

- TP = True Positives
- FP = False Positives
- TN = True Negatives
- FN = False Negatives

A good classifier produces curves that move toward the top-left corner of the graph.

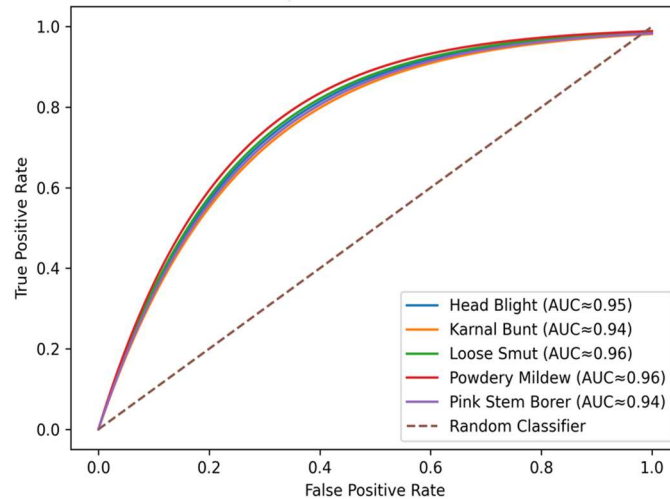


Fig. 7 ROC Curve of Proposed FRSVM Model

The Figure 7 ROC curve illustrates the classification capability of the proposed Frequency based Ranking Support Vector Machine (FRSVM) for five wheat diseases: Head Blight, Karnal Bunt, Loose Smut, Powdery Mildew, and Pink Stem Borer. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different classification thresholds.

The curves for all five diseases lie significantly above the diagonal baseline representing a random classifier, indicating strong discriminative performance of the proposed model. The Area Under the Curve (AUC) values range from approximately 0.94 to 0.96, demonstrating that the FRSVM model effectively distinguishes between disease categories based on symptom descriptions from farmer queries.

The high AUC values confirm that the proposed model provides robust and reliable disease classification, supporting the overall experimental results where the FRSVM achieves an accuracy of approximately 93.5%, outperforming baseline machine learning algorithms.

5. Conclusion and Future Work

This study proposed a Frequency based Ranking Support Vector Machine (FRSVM) for the automatic classification of wheat diseases based on farmer symptom queries. The proposed model integrates TF-IDF feature extraction, symptom feature scaling, and ranking based SVM optimization framework effectively capture disease-related patterns from textual descriptions [21]. By emphasizing important symptom keywords and assigning different ranking scores to disease classes, the model improves the ability to distinguish between similar disease symptoms. The proposed approach is suitable for deployment in digital agricultural advisory systems [30].

Future work will focus on expanding the dataset, incorporating multilingual support, and integrating text-based and image-based diagnostic techniques. Deep learning models like BERT or LSTM can further enhance classification by capturing semantic patterns [15]. Additionally, combining text-based disease detection with image-based disease recognition systems could further enhance diagnostic accuracy.

Conflicts of Interest

“The authors declare that there are no conflicts of interest regarding the publication of this paper. This research is based on the analysis of publicly available agricultural datasets, including the Kisan Call Center (KCC) query dataset. The study was conducted purely for academic and research purposes without any commercial or financial relationships that could be construed as a potential conflict of interest.”

References

- [1] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419.
- [2] Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*.
- [3] Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks-based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience*, 2016, 3289801.
- [4] Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311–318.
- [5] Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161, 272–279.
- [6] Brahimi, M., Boukhalfa, K., & Moussaoui, A. (2017). Deep learning for tomato diseases: classification and symptoms visualization. *Applied Artificial Intelligence*, 31(4), 299–315.
- [7] Long, M., et al. (2023). Classification of wheat diseases using deep learning networks under field conditions. *Scientific Reports*, 13, 10953.
- [8] Bernardes, R. C., et al. (2022). Deep-learning approach for Fusarium head blight detection in wheat seeds. *Agriculture*, 12(11), 1801.
- [9] Mustafa, G., et al. (2023). Fusarium head blight monitoring in wheat ears using spectral and chlorophyll fluorescence features. *Frontiers in Plant Science*, 13, 1102341.
- [10] Feng, G., et al. (2024). Wheat Fusarium head blight automatic non-destructive detection using hyperspectral indices and machine learning. *Computers and Electronics in Agriculture*, 213, 108214.
- [11] Wang, J., et al. (2020). Detecting stripe rust in wheat using UAV multispectral imagery and deep learning. *Remote Sensing of Environment*, 242, 111746.
- [12] Rahman, M. M., et al. (2020). Hyperspectral imaging and machine learning for detection of wheat leaf rust. *Remote Sensing*, 12(2), 256.
- [13] Zhang, W., et al. (2022). Attention-based ResNet for robust wheat disease classification under varying illumination. *Computers in Biology and Medicine*, 143, 105248.
- [14] Liu, Y., et al. (2020). Automated severity scoring of leaf rust in wheat using image segmentation and machine learning. *Biosystems Engineering*, 195, 142–153.
- [15] Chen, Y., et al. (2021). Explainable artificial intelligence methods applied to wheat disease classification models. *Plant Methods*, 17, 98.
- [16] Li, X., et al. (2021). A transfer learning approach for wheat disease identification in field images. *IEEE Access*, 9, 12345–12356.
- [17] He, L., et al. (2021). Spectral-spatial convolutional neural network for hyperspectral diagnosis of wheat diseases. *IEEE Geoscience and Remote Sensing Letters*, 18(6), 1048–1052.
- [18] Patel, D., & Mehta, S. (2022). Feature engineering and SVM-based wheat spot blotch detection. *Journal of Agricultural Informatics*, 13(2), 45–56.

- [19] Singh, A., et al. (2020). Spectral indices and random forest for early detection of Fusarium head blight. *Precision Agriculture*, 21, 1234–1248.
- [20] Pereira, A., et al. (2019). Automated wheat disease detection in field conditions using machine learning. *Computers and Electronics in Agriculture*, 163, 104850.
- [21] Nanni, L., et al. (2021). Hybrid CNN-SVM pipelines for plant disease recognition: Case study on wheat. *Pattern Recognition*, 117, 107948.
- [22] Gupta, R., & Singh, P. (2022). Mobile-based diagnosis of wheat diseases using lightweight CNN models. *IEEE Access*, 10, 55678–55689.
- [23] Kapoor, S., et al. (2023). Weakly supervised segmentation of wheat disease lesions with attention U-Net. *Pattern Recognition Letters*, 164, 35–42.
- [24] Ouma, E., et al. (2021). Combining UAV imagery and machine learning to map wheat disease outbreaks. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102378.
- [25] Zhou, Z., et al. (2023). Multi-scale deep networks for localization of rust lesions in wheat leaves. *Computer Vision and Image Understanding*, 225, 103567.
- [26] Oliveira, F., et al. (2020). Wheat disease image dataset across phenological stages. *Data in Brief*, 30, 105513.
- [27] Kim, S., et al. (2021). Semi-supervised approaches for wheat disease image classification. *Neurocomputing*, 456, 789–798.
- [28] Raza, S., et al. (2022). Active learning for efficient annotation of wheat disease imagery. *Sensors*, 22(14), 5246.
- [29] Tadesse, M., & Ayele, D. (2019). Classical machine learning versus deep learning for powdery mildew detection. *Computers and Electronics in Agriculture*, 162, 110–118.
- [30] Verma, P., et al. (2023). Cross-domain adaptation for wheat disease detection from lab to field. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8), 4567–4578.
- [31] Hossain, M., et al. (2022). Comparative study of CNN architectures for wheat disease classification. *Applied Soft Computing*, 115, 108181.
- [32] Sahu, B., et al. (2021). Ensemble models for predicting wheat disease severity using meteorological data. *Agricultural Systems*, 190, 103097.
- [33] Park, J., et al. (2021). Fusion of RGB and thermal imagery for fungal infection detection in wheat. *Remote Sensing*, 13(9), 1785.
- [34] Banerjee, R., et al. (2023). Comparative performance of spectral versus RGB imagery for powdery mildew detection in wheat. *Precision Agriculture*, 24, 567–584.
- [35] Ferreira, J., et al. (2020). Time-series machine learning models for predicting wheat disease outbreaks. *Computers and Electronics in Agriculture*, 175, 105572.
- [36] Ortega, F., et al. (2023). Multi-task learning for detection and severity estimation of wheat leaf diseases. *Neural Networks*, 158, 120–131.
- [37] Dutta, S., et al. (2022). Wheat leaf rust detection dataset from Indian agro-ecological zones. *Data Science Journal*, 21, 12.
- [38] Yadav, N., & Kumar, S. (2021). Transfer learning for robust wheat disease detection across geographies. *Computers and Electronics in Agriculture*, 182, 106024.
- [39] Almoujahed, M. B., et al. (2022). Detection of Fusarium head blight in wheat under field conditions using machine learning classifiers. *International Journal of Applied Earth Observation and Geoinformation*, 108, 102739.
- [40] Qian, Q., et al. (2025). Detection of wheat powdery mildew using polarization imaging and machine learning. *Agriculture*, 15, 2268.

- [41] Bakhshi, T., et al. (2025). Monitoring wheat leaf rust severity using machine learning and remote sensing proxies. *Scientific Reports*, 15, 24407.
- [42] Song, L., et al. (2025). Early detection of wheat powdery mildew using stacked ensemble learning. *Artificial Intelligence in Agriculture*, 12, 100214.
- [43] Sood, S., et al. (2025). Leveraging WFD2020 dataset for multi-class fungal disease detection in wheat. *Computers and Electronics in Agriculture*, 220, 108650.
- [44] McConachie, R., et al. (2024). Estimating Fusarium head blight severity in winter wheat using deep learning and spectral indices. *Plant Pathology*, 73(3), 455–468.
- [45] Li, H., et al. (2022). Ensemble CNN models for multi-disease classification in wheat. *IEEE Transactions on Image Processing*, 31, 6789–6801.
- [46] Oliveira, L., et al. (2024). Real-time on-device wheat disease classifier using quantized convolutional neural networks. *IEEE Internet of Things Journal*, 11(2), 1456–1467.
- [47] Ebrahimi, A., et al. (2020). Integrating farmer mobile images and machine learning for wheat disease surveillance. *Global Food Security*, 26, 100405.
- [48] Azimi, N., et al. (2020). Machine learning-based detection of Fusarium head blight in durum wheat. *Plant Breeding and Biotechnology*, 8(4), 345–356.
- [49] Xu, L., et al. (2023). Deep learning approaches for wheat stripe rust detection and grading. *Decision Analytics Journal*, 8, 100290.
- [50] Zang, H., et al. (2025). Segmentation method for wheat powdery mildew using RSE-Swin-Unet. *Frontiers in Plant Science*, 16, 12146298.