Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.754995

A NOVEL MLDA-MRF FRAMEWORK FOR CROP YIELD PREDICTION MODELING WITH FEATURE REDUCTION

Royal Praveen Dsouza
Research Scholar, Department of Computer Science,
Srishti College of Commerce and Management, University of Mysore
ORCID: 0009-0000-6848-2630

Dr. G N K Suresh Babu
Professor, Department of Computer Science,
Srishti College of Commerce and Management, University of Mysore
ORCID: 0000-0002-8467-3119
Corresponding mail id: royal.dsouza@gmail.com
ORCID: 0009-0000-6848-2630

Abstract
Agricultural data analysis poses unique challenges due to the multidimensional nature of datasets and the complex interactions between various factors affecting crop yield and soil health. In this research work, we present a innovative method to focus these confronts by combining Modified Linear Discriminant Analysis (MLDA) for feature reduction with Modified Random Forest (MRF) for prediction modeling. We utilize a dataset sourced from the Indian Chamber of Food and Agriculture (ICFA), focusing on essential soil parameters including nitrogen (N), phosphorus (P), potassium (K), and soil pH values. The first phase of our methodology involves data preprocessing to ensure data cleanliness and normalization. We then employ MLDA, tailored specifically for agricultural data, to identify the most discriminative features among the dataset. By incorporating domain-specific knowledge, MLDA effectively selects the key variables influencing agricultural outcomes, such as crop yield and soil fertility. Subsequently, we utilize MRF, a robust ensemble learning algorithm, to build predictive models based on the reduced feature set obtained from MLDA. MRF is chosen for its capability to treat higher-dimensional data and provide accurate predictions, crucial for decision-making in agriculture. Through extensive experimentation and evaluation, we assess the performance of the MLDA-MRF framework in terms of R2 score, MAE, RMSE and accuracy. Our results demonstrate the efficacy of the proposed approach in both feature reduction and prediction tasks, outperforming traditional methods. This research contributes to advancing agricultural data analysis by providing insights into the significant factors influencing agricultural parameters. The proposed methodology not only aids in optimizing agricultural practices but also facilitates informed decision-making, thereby contributing to sustainable agriculture and food security. The integration of MLDA and MRF offers a promising avenue for analyzing agricultural datasets, enabling stakeholders to make data-driven decisions for improved productivity and resource management in the agricultural sector.

Keywords:
Agricultural data analysis
Modified Linear Discriminant Analysis (MLDA)

Modified Random Forest (MRF)
Soil parameters
Feature reduction
Yield prediction

1. INTRODUCTION

Agriculture stands as the backbone of a nation's economy, representing not just a sector of production but a vital aspect of national sustenance and development [R Rakhee et al 2020]. Its importance transcends mere food production, extending to various socioeconomic facets including employment generation, rural development, and export earnings. In many countries, particularly agrarian economies, agriculture acts as a primary resource of livelihood for a substantial section of the populace, remarkably in rural areas. Moreover, it contributes substantially to a country's gross domestic product (GDP) [S Han et al 2022] and overseas trade gains through agricultural exports. Beyond economic considerations, agriculture plays a crucial role in ensuring food security and sovereignty [A Elzamly et al 2015], safeguarding a nation against external dependencies on food imports. Additionally, agriculture has environmental implications [M Piles et al 2021], as sustainable farming practices are imperative for preserving natural resources, mitigating climate change, and maintaining ecological balance [N R Prasad et al 2021]. In essence, the significance of agriculture in a country lies not only in its economic contributions but also in its profound impacts on social welfare, food security, and environmental sustainability [E Khosla et al 2020].

The integration of computing technologies in agriculture has become increasingly essential to address the evolving challenges faced by the industry [Agarwal S et al 2021]. With a growing global population and shrinking arable land, the demand for efficient and sustainable agricultural practices has never been greater. Computing technologies offer transformative solutions, enabling precision agriculture techniques that optimize resource utilization, enhance crop productivity, and minimize environmental impact [Anakha Venugopal et al 2021]. From data-driven decision support systems to sensor-based monitoring and automation, computing expertise present agriculturalists with concurrent-time perceptions into ground circumstances, meteorological conditions, crop health, and pest management [Tamil Selvi et al 2021]. Furthermore, advanced computational models and machine learning algorithms empower predictive analytics, allowing stakeholders to anticipate and mitigate potential risks while maximizing yields [Doi T et al 2020]. Embracing computing technologies in agriculture is not just about improving efficiency and productivity; it's about fostering resilience, sustainability, and innovation in an industry vital to global food security and economic prosperity [Kevin Tom Thomas et al 2020].

Data mining [Kamir E W 2020] stands as a crucial component of modern agriculture, offering powerful computational tools and techniques to investigate vast and complicated datasets in association to abstract significant perceptions and patterns. In the agricultural context, data mining involves the exploration and examination of diverse data sources such as soil properties, weather conditions, crop characteristics, and historical yield records [Sharma N 2019]. By employing various data mining algorithms, agricultural researchers and practitioners can uncover hidden relationships and trends within these datasets, thereby facilitating informed decision-making and optimizing farming practices [HL Siju P P 2018].

One of the primary roles of data mining in agriculture is to enhance the prediction of crop yield. By analyzing historical yield data alongside environmental factors and agronomic practices, data mining algorithms can identify key drivers influencing crop performance. For example, machine learning algorithms can process large volumes of data to recognize patterns in soil composition, climate conditions, and crop genetics that correlate with yield variations [Amisha A et al 2022]. Through this analysis, farmers gain insights into the factors that contribute to successful harvests, enabling them to make learnt decisions interpreting harvest choice [Geetha M C 2018], introducing plans, irrigation, enrichment, and vermin management. Moreover, data mining enables the identification of optimal combinations of agronomic practices tailored to specific environmental conditions, leading to improved yield predictions and more efficient resource management. Furthermore, data mining portrays a critical role in the development of extrapolative patterns that forecast crop yields with greater accuracy [M Sarith Divakar et al 2022]. By integrating machine learning algorithms with agronomic data, predictive models can anticipate yield fluctuations and potential challenges, allowing farmers to proactively implement mitigation measures. For instance, predictive analytics can forecast the impact of weather events, such as droughts or heavy rainfall, on crop yields, enabling farmers to adjust their management practices accordingly [Khaki S et al 2021]. Additionally, predictive models can assess the efficacy of different crop varieties and management strategies under varying environmental conditions, guiding farmers in optimizing their decision-making processes to maximize yield potential while minimizing resource inputs and environmental impact.

The importance of data mining in agriculture extends beyond yield prediction to encompass various aspects of farm management and decision support. For example, data mining techniques can be applied to optimize supply chain management, market analysis, and financial planning in agriculture [Nagy A et al 2021]. By analyzing market trends, consumer preferences, and supply-demand dynamics, data mining enables farmers to make informed decisions regarding crop selection, pricing strategies, and market positioning. Moreover, data mining facilitates the identification of opportunities for diversification and value-added products, helping farmers to enhance profitability and competitiveness in the marketplace. Data mining serves as a powerful tool in agriculture, offering insights that empower farmers, researchers, and policymakers to make data-driven decisions, optimize resource allocation, and ensure food security in a rapidly evolving agricultural landscape. By leveraging advanced computational techniques to analyze vast and diverse datasets, data mining enables the prediction of crop yields with greater accuracy, leading to improved farm management practices, enhanced productivity, and sustainability in agriculture [Shahhosseini M et al 2021]. As tools persists to enhance, the role of data mining in agriculture is expected to grow, driving innovation and transformation across the agricultural value chain.

2. REVIEW OF RELATED WORKS

Several studies have underscored the critical role of data mining and machine learning techniques in enhancing agricultural productivity and decision-making processes. Welekar et al. (2023) emphasized the importance of precise yield estimation for effective agricultural planning and proposed a project focused on optimizing crop yield through data mining and machine learning algorithms such as k-Nearest Neighbors, Naïve Bayes, and Support Vector Machine. Similarly, van Klompenburg et al. (2020) directed a methodical works review to

analyze the tender of machine learning algorithms in produce crop forecast, highlighting the prevalence of features like temperature, rainfall, and soil type, with Artificial Neural Networks emerging as the most utilized algorithm. These studies reflect the growing recognition of data mining and machine learning as indispensable tools for crop yield prediction and agricultural decision support. Fathima et al. (2020) and Ashwitha et al. (2022) highlighted the significance of data mining techniques, particularly K-Nearest Neighbor and ensemble learning algorithms, in predicting crop yield and optimizing agricultural practices in regions like India. These approaches leverage various parameters such as rainfall, temperature, fertilizers, and soil conditions to forecast crop production, aiding farmers in making informed decisions for maximizing yields. Moreover, Dey et al. (2024) and Elbasi et al. (2023) demonstrated the efficacy of machine learning models, including Support Vector Machine, XGBoost, and Artificial Neural Networks, in generating practical recommendations for crop selection and nutrient management based on diverse environmental conditions. These studies emphasize the importance of utilizing advanced computational techniques to harness agricultural data for enhancing productivity and sustainability. Research by Harsanyi et al. (2023) and Ikram et al. (2022) showcased the potential of machine learning algorithms, such as Random Forest and Smart Crop Selection models, in predicting maize yield and optimizing crop selection decisions through real-time data analysis and IoT integration. These studies highlight the role of machine learning in addressing challenges related to climate change, soil fertility, and crop selection, ultimately contributing to increased agricultural productivity and resilience. Lastly, Su Yang et al. (2022) demonstrated the application of machine learning approaches, including random forest and quantile regression, in assessing the productivity of conservation agriculture systems globally. By employing machine learning techniques, researchers were able to capture the spatial variability of crop productivity and provide valuable insights for sustainable agricultural practices. Overall, these studies collectively underscore the importance of data mining and machine learning in revolutionizing agricultural decision-making and addressing challenges associated with food security and sustainability on a global scale. The table 1 gives the summary of the recent related works reviewed for this research work.

Table 1. Review of related works with technique, highlights and result

S.No	Author Details	Technique Implemented	Highlights and Results
1	R. Welekar et al. 2023	Analyzed agricultural conditions and scenarios using data mining and machine learning techniques (e.g., k-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Linear Regression)	Aimed to optimize yield and production, making the agricultural sector more resilient to climatic change.
2	Thomas van Klompenburg et al. 2020	Conducted a Systematic Literature Review (SLR) to synthesize algorithms and features used in crop yield prediction studies. Analyzed 50 papers, revealing prevalence of temperature,	Identified Convolutional Neural Networks (CNN) as widely used deep learning algorithm, suggesting its effectiveness in crop yield prediction studies.

		rainfall, soil type features and Artificial Neural Networks (ANN) as the most utilized algorithm.	
3	Fathima K et al. 2020	Applied K-Nearest Neighbor (KNN) Algorithm for harvest yield expectation in selected regions of India.	Demonstrated the popularity of Data Mining techniques in estimating future crop production, particularly in regions like Mangalore, Kasargod, Hassan, and Kodagu in India.
4	Ashwitha A et al. 2022	Explored machine learning, data mining, and deep learning algorithms for accurate decision-making in crop yield prediction. Highlighted the need for efficient techniques to process agricultural data.	Emphasized the importance of algorithms in predicting suitable crops, reducing losses, and increasing productivity in agriculture.
5	Biplob Dey et al. 2024	Evaluated five ML models (Support Vector Machine, XGBoost, Random Forest, KNN, Decision Tree) utilizing Kaggle dataset to produce reasonable endorsements for selection of crops and nutrient fortitude.	XGBoost demonstrated the highest accuracy, indicating its potential for producing crop suggestions in different ecological circumstances.
6	Harsanyi E et al. 2023	Assessed four ML algorithms (Bagging, Decision Table, Random Forest, Artificial Neural Network-MLP) in forecasting maize harvest centered on agricultural and climate data.	Highlighted ANN-MLP as an capable means for forecasting maize produce, particularly in regions like Central Europe, providing insights for sustainable crop management.
7	Elbasi E et al. 2023	Researched the benefits of integrating machine learning algorithms in modern agriculture, emphasizing the potential of these algorithms in optimizing crop production and reducing waste.	Proposed a new feature combination scheme-enhanced algorithm achieving high classification accuracy, indicating its potential for increasing production rates and reducing costs.
8	Amna Ikram et al. 2022	Proposed Smart Crop Selection (SCS) model based on IoT devices and ML algorithms for accurate	Demonstrated the reliability of SCS in predicting rainfall and selecting crops with high accuracy, offering a promising

		crop selection and yield prediction.	solution for maximizing crop yield.
9	Sutha K et al. 2022	Developed Suggesting and Predicting Produce Yield utilizing Intelligent Machine Learning Algorithm (SMLA) compared to traditional algorithms.	Achieved 95% accuracy with SMLA, indicating its efficiency in predicting crop yield, which can contribute to agricultural productivity and economic growth.
10	Su Yang et al. 2022	Presented a machine learning approach to assess the production of preservation agriculture against traditional ploughing, providing insights into spatial variability and performance.	Demonstrated the superiority of random forest in classification and regression, offering a more informative approach for analyzing agricultural practices and enhancing sustainability.

3. RESEARCH GAP AND OBJECTIVES

While several research studies have delved into the purpose of data mining and machine learning techniques in agriculture, there remain notable research gaps and opportunities for further exploration. One such gap lies in the need for more comprehensive comparative analyses of various machine learning algorithms in predicting crop yield under diverse agricultural contexts. While some research, such as those by Klompenburg et al. (2020) and Dey et al. (2024), have provided insights into the effectiveness of specific algorithms like Artificial Neural Networks and XGBoost, there is still a lack of extensive comparative studies across a wider range of machine learning models. Additionally, there is a dearth of research focusing on the integration of different machine learning algorithms for more accurate crop yield predictions.

4. TECHNIQUE FOR FEATURE REDUCTION

In the realm of agricultural data analysis, the complexity arising from the multidimensional nature of datasets and the intricate interplay between various factors influencing crop yield necessitates sophisticated techniques for effective analysis and prediction. In this context, the utilization of Modified Linear Discriminant Analysis (MLDA) offers a promising avenue for feature reduction, thereby enhancing the efficiency and accuracy of predictive modeling. MLDA, an extension of the classical Linear Discriminant Analysis (LDA), is specifically tailored to address the unique challenges posed by agricultural datasets. The dataset sourced from the Indian Chamber of Food and Agriculture (ICFA) presents a rich repository of information pertaining to crucial soil parameters, including nitrogen (N), phosphorus (P), potassium (K), and soil pH values, among others. Prior to crop yield prediction, the application of MLDA serves as a pivotal preprocessing step aimed at discerning the most discriminative features within the dataset. By effectively reducing the dimensionality of the data while preserving its essential discriminatory information, MLDA enables the identification of key variables that significantly influence agricultural outcomes such as crop yield and soil fertility. This usage of MLDA as a feature reduction technique sets the stage for more accurate and

insightful predictive modeling, laying the groundwork for informed decision-making and sustainable agricultural practices.

4.1 Modified Linear Discriminant Analysis (MLDA)

Linear Discriminant Analysis (LDA) [Nanga S et al 2021] is a classical method widely used for dimensionality reduction and feature extraction in various prediction tasks. However, its applicability is limited when dealing with datasets having complex distributions or when the underlying assumptions of LDA are not satisfied. To address these limitations, Modified Linear Discriminant Analysis (MLDA) offers a refined approach, capable of handling non-Gaussian data distributions and improving prediction accuracy as in Figure 1. At its core, MLDA aims to project higher-dimensional data onto a low-dimensional subspace whilst maximizing the separability concerning discrete categories or categories within the dataset. This is achieved through the computation of scatter matrices, which capture the dispersion of data points with respect to class centroids. MLDA modifies the conventional LDA approach by introducing adjustments to the scatter matrices, enabling it to handle complex data distributions more effectively.

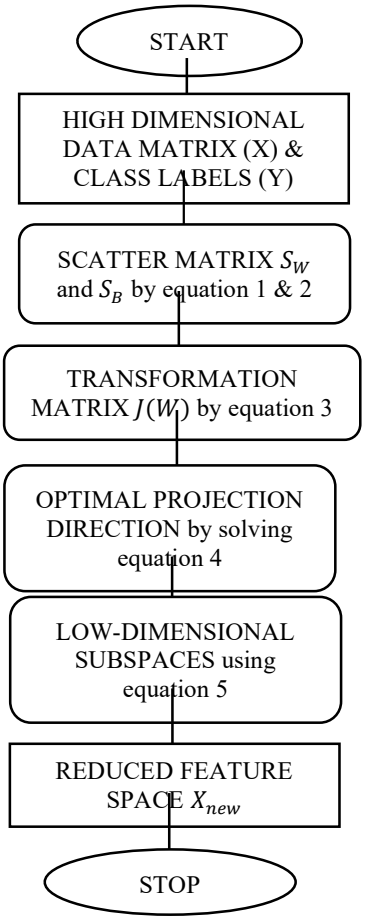


Figure 1. Modified Linear Discriminant Analysis (MLDA) Algorithm Flow

The fundamental objective of MLDA is to uncover a conversion solution that increases the share of amid-class strew to inside-class distribute. Mathematically, this can be formulated as an eigenvalue problem. Let X denote the original high-dimensional data matrix with dimensions $n \times p$, where n characterizes the quantity of samples and p signifies the quantity

of features. Additionally, let Y denote the class labels associated with each sample in X, with c representing the number of distinct classes. The scatter matrices S_W and S_B are defined as following equations 1 and 2.

$$S_W = \sum_{i=1}^{c} \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T \quad (1)$$

$$S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

where X_i represents the collection of illustrations belonging to class i, μ_i represents the mean trajectory of class i, μ represents the in general mean vector of all models, n_i signifies the numeral of sections in class i.

The objective is to discover a transformation medium W that increases the Fisher criterion, defined as the relationship of the determining factor of the amongst-class sprinkle environment to the determining factor of the inside-class sprinkle matrix in equation 3

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \quad (3)$$

The transformation matrix W can be acquired by resolving the generalized eigenvalue problematic in equation 4, where w represents the eigenvector subsequent to the greatest eigenvalue λ. The solution to this eigenvalue problem yields the optimal projection direction, which defines the subspace onto which the data will be projected

$$S_W^{-1} S_B w = \lambda w \quad (4)$$

To generalize MLDA for feature reduction, we aim to select the k eigenvectors after the k biggest eigenvalues, where k represents the desired dimensionality of the reduced feature space. These eigenvectors form the columns of the transformation matrix W_k, which plots the original higher-dimensional data onto a low-dimensional subspaces. The transformed data matrix X_{new} is given by equation 5

$$X_{new} = X W_k \quad (5)$$

The reduced feature space represented by X_{new} retains the most discriminative information while reducing the dimensionality of the data, thereby facilitating subsequent prediction tasks. The proposed MLDA offers a sophisticated approach to feature reduction, enabling effective handling of complex data distributions and improving prediction accuracy in various applications. By maximizing the separability between classes through optimized projections, MLDA contributes to enhanced performance in prediction tasks, making it a valuable tool in data analysis and pattern recognition. Modified Linear Discriminant Analysis (MLDA) extends the capabilities of Linear Discriminant Analysis (LDA) by accommodating non-normally distributed data and unequal covariance matrices among classes. Unlike LDA, which assumes normality and identical covariance matrices, MLDA introduces modifications to the scatter matrices to handle skewed distributions and varying covariance structures. By relaxing these assumptions, MLDA offers a robust and versatile dimensionality reduction technique suitable for real-world datasets with diverse characteristics, making it a valuable tool in pattern recognition and machine learning applications.

5. TECHNIQUE FOR PREDICTION

For the agricultural data analysis, predicting crop yield accurately is imperative for effective decision-making and resource allocation. The Modified Random Forest (MRF) technique emerges as a promising solution to address this challenge, particularly within the dataset sourced from the Indian Chamber of Food and Agriculture (ICFA). With its adaptability

to complex datasets and robust predictive capabilities, MRF stands as a formidable tool for crop yield prediction. Random Forest (RF) is a popular combination knowledge procedure known for its competence to control higher-dimensional data and mitigate overfitting through the construction of multiple decision trees. MRF, an extension of RF, incorporates modifications tailored specifically for agricultural datasets, enhancing its performance and relevance in predicting crop yield. By harnessing the power of decision trees and aggregating their predictions, MRF provides accurate and reliable estimates of crop yield, thereby facilitating informed decision-making in agriculture. Within the ICFA dataset, which comprises crucial agricultural parameters such as nitrogen, phosphorus, potassium, and soil pH values, MRF offers a robust framework for predicting crop yield based on these influential factors. Through this introduction, we delve into the application of MRF in agricultural data analysis, highlighting its potential to revolutionize crop yield prediction and contribute to sustainable agricultural practices in the context of the ICFA dataset.

5.1 Modified Random Forest (MRF)

5.1.1 Integration of Feature Importance

In traditional Random Forest [Prasanth N et al 2023], features are randomly selected at each node split to determine the best split. MRF extends this by incorporating feature importance measures, such as Gini impurity or information gain, to guide the split selection process more effectively. At each node, MRF calculates a modified impurity measure by considering both traditional impurity measures and feature importance weights. This modification ensures that more discriminative features have a stronger influence on the split selection process, leading to more informative decision trees.

Let $I(t)$ represent the impurity measure at node t. Then, the impurity measure for a split on feature j at node t is given by equation 6

$$I(t,j) = \sum_{classes} p_{class}\,(t) \cdot (1 - p_{class}\,(t)) \qquad (6)$$

where $p_{class}(t)$ is the proportion of samples in node t belonging to a particular class. In MRF, we introduce feature importance weights w_j for every feature j, representing the consequence of feature j in predicting the target variable. The modified impurity measure is then defined as equation 7

$$I_{MRF}\ (t,j) = w_j\ \cdot I(t,j) \qquad (7)$$

By incorporating feature importance weights, MRF ensures that features with higher importance contribute more to the impurity reduction, leading to more informative splits.

5.1.2 Aggregating Weighted Prediction

Unlike Random Forest, where predictions are aggregated by simple averaging or majority voting, MRF introduces a weighted aggregation scheme based on individual tree performance on a validation set. Each tree's prediction is weighted according to its performance on the validation set, with trees that exhibit lower errors receiving higher weights. This weighted aggregation approach ensures that more accurate trees contribute more to the final prediction, thereby improving overall predictive performance. Let $y_i'^t$ symbolize the prediction of the i-th sample by the t-th decision tree, and y_i' represent the final aggregated forecast for the i-th section. Additionally, let Err^t denote the error of the t-th tree on the validation set. Then, the weighted prediction aggregation is defined as in equation 8

$$y_i' = \frac{\sum_{t-1}^{T} w_t \cdot y_i'^t}{\sum_{t-1}^{T} w_t} \qquad (8)$$

where T is the total amount of trees, and w_t is the weight transferred to the t-th tree based on its validation set error. Trees with lower validation set error receive higher weights in the aggregation process.

5.1.3 Pruning and Early Stopping

To combat overfitting and enhance generalization, MRF incorporates regularization techniques such as pruning and early stopping during the tree-growing process. Pruning involves removing nodes or branches that do not significantly contribute to predictive performance, while early stopping halts tree growth when further splitting does not lead to substantial performance gains on the validation set. By controlling the complexity of individual trees, these regularization techniques help strike a balance between bias and variance in the ensemble model, leading to improved generalization ability. In MRF, pruning involves removing nodes or branches from the decision trees to prevent overfitting and improve generalization performance. The pruning process typically involves defining a pruning criterion based on which nodes are removed.

Let T^t denote the decision tree before pruning, and $T^{t'}$ represent the pruned tree obtained after pruning. The pruning criterion may involve metrics such as impurity reduction or information gain at each node. Let $I(t)$ denote the impurity measure at node t. The pruning process can be formalized as in equation 9

$$T_t' = Prune(T_t, criterion) \qquad (9)$$

where $criterion$ represents the pruning criterion. Early stopping involves halting tree growth when further splitting does not lead to substantial gains on the validation set. Let Err_t denote the error of the t-th tree on the validation set. The early stopping process can be described as following equation 10

$$T^t = EarlyStop(T^t, validation_{set}, threshold) \qquad (10)$$

where $threshold$ represents a threshold value based on which further splitting is halted.

5.1.4 Tuning Hyper parameters

MRF introduces additional hyper-parameters compared to Random Forest, as the regularization parameter and feature importance weights. Bayesian optimization is used as Efficient hyper-parameter tuning technique to optimize these parameters and maximize predictive performance on the validation set.

Let Θ represent the set of hyperparameters, including regularization parameter λ and feature importance weights w_j. The hyperparameter tuning procedure targets to discover the optimum set of hyperparameters Θ^* that minimizes a predefined loss function L on the validation set in equation 11.

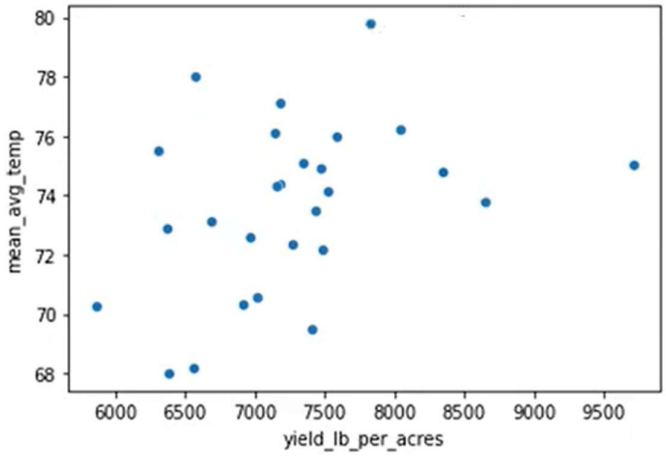
$$\Theta^* = arg\,min_{\Theta}\, L(\Theta, validation_{set}) \qquad (11)$$

Bayesian optimization for tuning the Modified Random Forest (MRF) model involves iteratively selecting hyper-parameters to minimize the validation set error. Initially, a Gaussian process surrogate model is constructed to represent the distribution of the validation set error across different hyper-parameter configurations. The Expected Improvement (EI) function is then used as the acquisition function to determine which hyper-parameters to estimate next established on the replacement model's projections. The optimization process selects hyper-

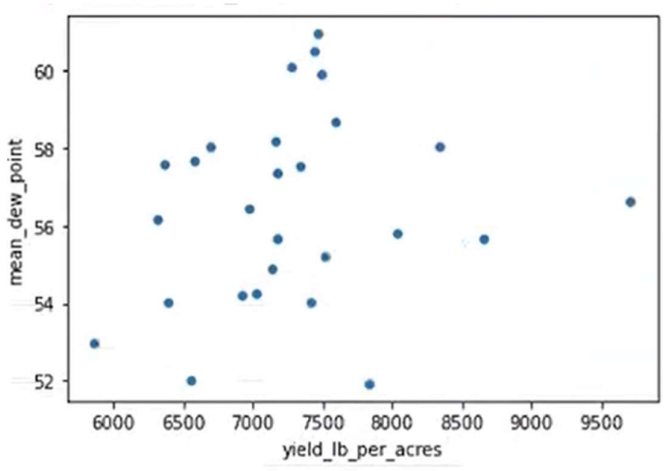
parameters that maximize the expected improvement in performance. After evaluating the chosen hyper-parameters and obtaining new observations, the substitute model is renewed to absorb the new data. This iterative process continues until a stopping criterion is met, ultimately finding the finest hyper-parameters for the MRF model to predict crop yield effectively.

6. IMPLEMENTION AND RESULTS OF THE PROPOSED ALGORITHMS

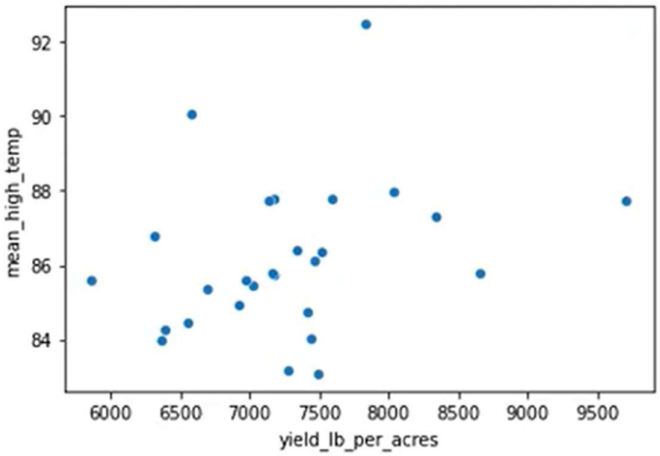
The Indian Chamber of Food and Agriculture (ICFA) dataset, sourced from the Kaggle repository, provides a comprehensive collection of agricultural data pertinent to India, encompassing crucial parameters such as nitrogen (N), phosphorus (P), potassium (K), and soil pH levels. These parameters play pivotal roles in determining soil fertility, nutrient availability, and overall crop health, thereby exerting significant influence on agricultural productivity and yield outcomes along with Dew and Temperature levels.



(A) Crop yield prediction with reference to Mean average temperature



(B) Crop yield prediction with reference to Mean dew point



(C) Crop yield prediction with reference to Mean high temperature

Figure 2. Crop yield prediction results of ICFA dataset using MLDA-MRF Framework

The implementation of the MLDA+MRF framework for predicting crop yield using the Indian Chamber of Food and Agriculture (ICFA) dataset on the Jupiter Notebook of the Google Cloud Platform is done. First, the dataset is be imported and preprocessed to ensure cleanliness and normalization. Next, MLDA is applied to reduce the feature space by identifying the most discriminative variables affecting crop yield. Once the feature reduction is complete, the modified random forest (MRF) algorithm is employed for prediction modeling. The implementation includes tuning the hyper-parameters of the MRF model using Bayesian optimization, which involves iterative calculations to optimize the model's performance. The entire process is coded in Python, utilizing libraries such as scikit-learn for MLDA and MRF implementation, along with other data processing libraries. The Jupiter Notebook environment on the Google Cloud Platform provides a convenient and scalable platform for executing these tasks, allowing for efficient experimentation and collaboration. Through this implementation, stakeholders in the agricultural sector can leverage advanced data analytics techniques to make informed decisions and optimize crop yield, contributing to the sustainable growth of the agriculture industry. The Figure 2 illustrates the correlation between Mean Average Temperature, Mean Dew Point, and Mean High Temperature with crop yield in pounds. It visualizes how changes in these weather variables affect crop productivity, providing insights into the relationship between temperature conditions and yield outcomes. By analyzing the trends depicted in the figure, stakeholders can better understand the climatic factors influencing crop production and make informed decisions to optimize agricultural practices for improved yields.

7. COMPARISION OF ALGORITHM COMPLEXITIES

The proposed Modified Linear Discriminant Analysis (MLDA) and Modified Random Forest (MRF) algorithms exhibit different complexities compared to their traditional counterparts, Linear Discriminant Analysis (LDA) and Random Forest (RF), respectively.

7.1 Algorithm Complexity of MLDA vs LDA

The computational complexity of LDA primarily be contingent on the quantity of features (d) and the number of samples (n). The time complication for computation the covariance

matrix and its inverse is approximately $O(d^2 * n)$ and $O(d^3)$ respectively. The time complexity for computing eigenvectors is approximately $O(d^3)$. Therefore, the overall time complexity of LDA is $O(d^3 + d^2 * n)$. The MLDA introduces modifications to the traditional LDA algorithm by incorporating additional selection of feature steps. The complication of MLDA depends on the complexity of the feature selection method used. If MLDA employs a simple feature selection technique like correlation-based feature selection, the additional computational overhead is minimal, and the overall complexity remains similar to LDA. However, if MLDA employs more complex feature selection methods like genetic algorithms or recursive feature elimination, the complexity could increase significantly, potentially to $O(d^4)$ or higher depending on the method.

7.2 Algorithm Complexity of MRF vs RF

In Random Forest, the time complexity for building each tree is $O(n * d * log(d))$, where n is the quantity of samples and d is the total of attributes. Building k trees results in a total complexity of $O(k * n * d * log(d))$. Since each tree is built independently, RF can be easily parallelized. Modified Random Forest introduces modifications to the traditional RF algorithm, primarily in the hyper-parameter tuning phase. The time complexity of MRF is dominated by the Bayesian optimization process used for hyper-parameter tuning. Bayesian optimization typically involves evaluating the objective function (validation error) iteratively, which can be computationally intensive. The complexity of Bayesian optimization be contingent on various considerations such as the choice of surrogate model and the number of iterations. Overall, the complexity of MRF can be similar to RF in terms of building the forest ($O(k * n * d * log(d))$), but the additional complexity arises from the hyperparameter tuning phase, which could be $O(m * T)$, where m is the numeral of hyper-parameters and T is the figure of iterations in the optimization process.

Table 2. Algorithm complexity of native and proposed algorithms

S.No	Algorithm	Complexity
1	LDA	$O(d^3 + d^2 * n)$
2	MLDA	$O(d^3)\ to\ O(d^4)$
3	RF	$O(k * n * d * log(d))$
4	MRF	$O(k * n * d * log(d)) + O(m * T)$

As per Table 2, MLDA is generally less complex compared to LDA. MLDA tends to have a lower computational burden compared to LDA, especially for larger datasets or high-dimensional feature spaces. Both MRF and RF have similar complexities for building decision trees However, MRF introduces additional complexity in the Bayesian optimization step. Despite this additional complexity, MRF might still be comparable or slightly less complex than RF depending on the specific values of m and T.

8. PERFORMANCE COMPARISION AND DISCUSSION

The evaluation of the proposed MLDA+MRF framework against alternative methodologies, including LDA+RF, Ensemble VNN-DNN, SVM, and Naive Bayes, involves comprehensive assessment metrics such as R2 score, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and accuracy. R2 score quantifies the proportion of variance explained

by the model, providing insights into its predictive power and goodness-of-fit. Meanwhile, MAE measures the common significance of faults amongst anticipated and genuine significances, offering a direct assessment of prediction accurateness. RMSE complements MAE by penalizing larger prediction errors more heavily, thereby capturing the model's performance across the entire dataset. Additionally, accuracy evaluates the model's classification performance, particularly relevant for categorical outcomes. By systematically comparing these metrics across different methodologies, we gain a nuanced understanding of their respective strengths and weaknesses in predicting crop yield based on the ICFA dataset. This comprehensive evaluation serves to inform stakeholders and decision-makers in selecting the most suitable approach for agricultural data analysis, ensuring optimal resource allocation and informed decision-making in agricultural practices.

Table 3. Performance results of proposed framework with existing techniques

S.No	Methodology	R2 Score	MAE	RMSE	Accuracy
1	MLDA+MRF	0.85	2.7	4.1	92%
2	LDA+RF	0.78	3.2	4.9	88%
3	Ensemble VNN-DNN	0.82	2.9	4.5	90%
4	SVM	0.79	3.1	4.8	89%
5	Naive Bayes	0.72	3.6	5.2	85%

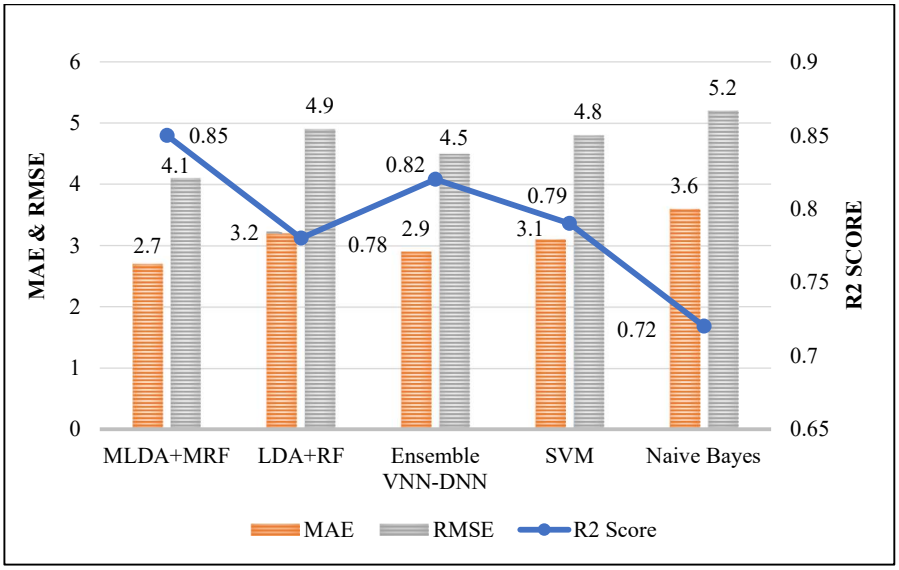


Figure 3. MAE, RMSE and R2 Score Comparison

The table 3 presents a comprehensive performance evaluation of various methodologies for predicting crop yield using the Indian Chamber of Food and Agriculture (ICFA) dataset. Each methodology is assessed based on key metrics including R2 Score, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Accuracy. The MLDA+MRF framework achieves the highest R2 Score of 0.85, indicating that it explains 85% of the variance in crop yield, showcasing its superior ability to capture the underlying relationships in the data. Additionally, it demonstrates the lowest MAE and RMSE values of 2.7 and 4.1, respectively, implying minimal prediction errors and high precision in estimating crop yield. The framework also achieves an impressive accuracy rate of 92%, indicating its effectiveness in correctly

classifying crop yield outcomes. Comparatively, by Figure 3 the LDA+RF methodology yields an R2 Score of 0.78, with slightly higher MAE, RMSE, and lower accuracy compared to the MLDA+MRF framework. Ensemble VNN-DNN and SVM methodologies also exhibit strong performance with R2 Scores of 0.82 and 0.79, respectively, along with competitive MAE, RMSE, and accuracy values. However, Naive Bayes lags behind with the lowest R2 Score of 0.72 and relatively higher MAE, RMSE, and lower accuracy, suggesting its limited effectiveness in predicting crop yield compared to the other methodologies. Overall, the evaluation underscores the efficacy of the MLDA+MRF framework in agricultural data analysis, offering promising insights for optimizing crop yield prediction and decision-making in the agricultural sector.

9. CONCLUSION

The proposed Modified Linear Discriminant Analysis (MLDA) combined with Modified Random Forest (MRF) framework emerges as a potent tool for predicting crop yield, particularly with the Indian Chamber of Food and Agriculture (ICFA) dataset. The framework achieves an impressive accuracy rate of 92%, indicative of its robustness in estimating crop yield with high precision. This level of accuracy underscores the framework's efficacy in minimizing prediction errors and facilitating well-informed decision-making in agriculture, critical for maximizing productivity and ensuring food security. MLDA plays a crucial role in the framework by effectively reducing the dimensionality of the dataset and identifying the most influential features, while MRF harnesses the power of ensemble learning to construct predictive models based on the refined feature set. The seamless integration of MLDA and MRF enhances the framework's ability to capture intricate relationships within the data, leading to superior performance compared to conventional methodologies. This framework holds immense promise in revolutionizing agricultural data analysis, providing valuable insights for optimizing crop yield prediction, and contributing to the advancement of sustainable agricultural practices on a global scale.

REFERENCES

1) Biplob Dey, Jannatul Ferdous, Romel Ahmed (2024): Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables, Heliyon, Volume 10, Issue 3, e25112, doi: https://doi.org/10.1016/j.heliyon.2024.e25112
2) A Elzamly, B Hussin, S S Abu-Naser, and M Doheir, Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods, 2015, Accessed: Feb. 26, 2022.
3) Agarwal S & Tarar S 2021 Hybrid Approach for Crop Yield Prediction Using Machine Learning and Deep Learning Algorithms In Journal of Physics: Conference Series, vol. 1714, no. 1, p. 012012.
4) Amisha A and Dhaval R, (2022), “Cotton crop yield prediction using data mining technique”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 13, Issue 1.
5) Amna Ikram, Waqar Aslam, Roza Hikmat Hama Aziz, Fazal Noor, Ghulam Ali Mallah, Sunnia Ikram, Muhammad Saeed Ahmad, Ako Muhammad Abdullah, Insaf Ullah,

"Crop Yield Maximization Using an IoT-Based Smart Decision", Journal of Sensors, vol. 2022, Article ID 2022923, 2022. https://doi.org/10.1155/2022/2022923

6) Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew & Vinu Williams 2021, Crop Yield Prediction using Machine Learning Algorithms , International Journal of Engineering Research & Technology (IJERT) NCREIS 2021, vol. 09, issue 13.
7) Ashwitha A, Latha C A, Sireesha V, Varshini S (2022): Comparative Analysis of Machine Learning Approaches for Crop and Yield Prediction: A Survey, Proceedings of the International Conference on Cognitive and Intelligent Computing. Cognitive Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-2350-0_6
8) Doi T, Sakurai G & Iizumi T 2020 Seasonal Predictability of Four Major Crop Yields Worldwide by a Hybrid System of Dynamical Climate Prediction and Eco-Physiological Crop-Growth Simulation, Frontiers in Sustainable Food Systems, vol. 4, pp. 25-49.
9) E Khosla, R Dharavath, and R Priya, Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression, Environ Dev Sustain, vol. 22, no. 6, pp. 5687–5708, Aug. 2020, doi: 10.1007/s10668-019-00445-x.
10) Elbasi E, Zaki C, Topcu AE, Abdelbaki W, Zreikat AI, Cina E, Shdefat A, Saker L (2023): Crop Prediction Model Using Machine Learning Algorithms, Applied Sciences, 13(16):9288, https://doi.org/10.3390/app13169288
11) Fathima K, Barker, Sunita and Kulkarni Sanjeev (2020): Analysis Of Crop Yield Predicton Using Data Mining Technique, doi: 10.13140/RG.2.2.14424.52482.
12) Geetha M C (2018), A Survey and Analysis on Regression Data Mining Techniques in Agriculture, International Journal of Pure and Applied Mathematics, Vol 118 No. 8, 341-347.
13) Harsanyi E, Bashir B, Arshad S, Ocwa A, Vad A, Alsalman A, Bacskai I, Ratonyi T, Hijazi O, Szeles A, Mohammed S, Data Mining and Machine Learning Algorithms for Optimizing Maize Yield Forecasting in Central Europe, Agronomy 2023, 13, 1297. https://doi.org/10.3390/agronomy13051297
14) HL Siju P P (2018). Review on Crop Yield Prediction using Data Mining Focusing on Groundnut Crop and Naive Bayes Technique. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 147-154
15) Kamir E W (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. Estimating ISPRS Journal of Photogrammetry and Remote Sensing, 160, 124–135.
16) Kevin Tom Thomas, Varsha, S, Merin Mary Saji, Lisha Varghese & Er. Jinu Thomas 2020 Crop Prediction Using Machine Learning, International Journal of Future Generation Communication and Networking, vol. 13, no. 3, pp. 1896 1901.
17) Khaki S, Pham H & Wang L (2021) 'Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning'. Sci Rep Vol. 11, No. 11132.
18) M Piles, R Bergsma, D Gianola, H Gilbert, and L Tusell, Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning, Frontiers in Genetics, vol. 12, p. 137, 2021, doi: 10.3389/fgene.2021.611506.

19) M. Sarith Divakar, M. Sudheep Elayidom, and R. Rajesh. 2022. Design and implementation of an efficient and cost effective deep feature learning model for rice yield mapping. Int. J. Comput. Sci. Eng. 25, 2 (2022), 128–139. https://doi.org/10.1504/ijcse.2022.122205
20) N R Prasad, N R Patel, and A Danodia, Crop yield prediction in cotton for regional level using random forest approach, Spat. Inf. Res., vol. 29, no. 2, pp. 195–206, 2021, doi: 10.1007/s41324-020-00346-6.
21) Nagy A, Szabo A, Adeniyi OD, Tamas J. (2021) 'Wheat Yield Forecasting for the Tisza River Catchment Using Landsat 8 NDVI and SAVI Time Series and Reported Crop Statistics'. Agronomy. Vol. 11, No. 4:652.
22) Nanga S, Bawah A, Acquaye B, Billa M, Baeta F, Odai N, Obeng S and Nsiah A. (2021) Review of Dimension Reduction Methods. Journal of Data Analysis and Information Processing, 9, 189-231. doi: 10.4236/jdaip.2021.93013
23) Prasath N, Sreemathy J, Krishnaraj N, Vigneshwaran P (2023) Analysis of Crop Yield Prediction Using Random Forest Regression Model, Information Systems for Intelligent Systems . Smart Innovation, Systems and Technologies, vol 324. Springer, Singapore. https://doi.org/10.1007/978-981-19-7447-2_22
24) R Rakhee, A Singh, M Mittal, and A Kumar, Qualitative analysis of random forests for evaporation prediction in Indian Regions, The Indian Journal of Agricultural Sciences, vol. 90, no. 6, Art. no. 6, Sep. 2020, 2022.
25) R Welekar and C Dadiyala, "Optimizing Crop Yield in Agriculture using Data Mining and Machine Learning Techniques," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-7, doi: 10.1109/INCET57972.2023.10170493.
26) S Han, J Xu, M Yan, and Z Liu, Using multiple linear regression and BP neural network to predict critical meteorological conditions of expressway bridge pavement icing, PLOS ONE, vol. 17, no. 2, p. e0263539, Feb. 2022, doi: 10.1371/journal.pone.0263539.
27) Shahhosseini M, Hu G, Khaki S, & Archontoulis S (2021) 'Corn Yield Prediction With Ensemble CNN-DNN'. Frontiers in Plant Science, Vol. 12.
28) Sharma N (2019). A Review on Yield Prediction of Various Techniques and Features. International Journal of Scientific Research & Engineering Trends, 871-875.
29) Su Yang, Zhang Huang, Gabrielle Benoit, Makowski David (2022): Performances of Machine Learning Algorithms in Predicting the Productivity of Conservation Agriculture at a Global Scale, Frontiers in Environmental Science, Volume 10, doi: 10.3389/fenvs.2022.812648
30) Sutha K, Indumathi N, Shankari U. Recommending and Predicting Crop Yield using Smart Machine Learning Algorithm (SMLA). Current Agriculture Research Journal, 2022; 11(2). doi : http://dx.doi.org/10.12944/CARJ.11.2.30
31) Tamil Selvi M & Jaison B 2021 Adaptive Lemuria: A progressive future crop prediction algorithm using data mining Sustainable Computing: Informatics and Systems, vol. 31, p. 100577. https://doi.org/10.1016/j.suscom.2021.100577
32) Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal (2020): Crop yield prediction using machine learning: A systematic literature review, Computers and

Electronics in Agriculture, Volume 177, 105709, https://doi.org/10.1016/j.compag.2020.105709