Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Optimizing Crop Yield Prediction - A Comparative Analysis and Development of a Hybrid Algorithm for Regional Agricultural Data

Bhavana Gowda D. M.[1], K. S. Arvind[2], Nachappa M. N[3]
[1]Research Scholar
School of Engineering and Technology
Jain (Deemed-to-be University), Bengaluru, India
Email: bgowdadm@gmail.com
[2]Associate Professor
Jain (Deemed-to-be University), Bengaluru, India
Email: ks.arvind@jainuniversity.ac.in
[3]Professor
School of Computer Science and Information Technology
Jain (Deemed-to-be University), Bengaluru, India
Email: mn.nachappa@jainuniversity.ac.in

ARTICLE INFO

Received: 20 Nov 2024

Revised: 25 Dec 2024

Accepted: 20 Jan 2025

ABSTRACT

Accurate crop yield prediction is essential for optimizing resource allocation, managing risks, and ensuring sustainable agricultural practices. This study introduces a novel hybrid algorithm that integrates multiple predictive models, including Linear Regression, Decision Trees, Random Forests, and Neural Networks, with a Gradient Boosting Machine (GBM) as the meta-model, to improve the accuracy of region-specific crop yield predictions. Using required dataset, covering environmental, agricultural, and economic factors, the hybrid algorithm demonstrated superior performance compared to individual models. It achieved an RMSE of 17.55 tons/ha, MAE of 13.80 tons/ha, and an R^2 of 0.87, outperforming state-of-the-art models. The study's findings underscore the hybrid algorithm's ability to capture complex, non-linear relationships in agricultural data, improving the precision of crop yield forecasts. This enhanced predictive capability can support farmers and policymakers in making informed decisions, optimizing resource use, and mitigating risks associated with climate variability. However, limitations such as dataset specificity and increased computational complexity highlight the need for further refinement. Future research should focus on expanding the dataset to diverse geographical regions and optimizing the algorithm for broader applicability.

Keyword: variability, algorithm, decisions, demonstrated, specificity

1. INTRODUCTION

Crop yield prediction is an essential aspect of modern agriculture, playing a pivotal role in ensuring food security, optimizing resource management, and improving farm operations. As the global population continues to expand, the demand for agricultural products is increasing, driving the need for more efficient and accurate methods of predicting crop yields. These predictions enable farmers and policymakers to make informed decisions regarding crop selection, planting schedules, resource allocation, and market planning, all of which contribute to enhanced economic outcomes and the sustainability of agricultural practices [1].

However, the accuracy of crop yield prediction models faces several challenges. Agricultural data is inherently heterogeneous, involving numerous factors such as soil properties, weather conditions, crop types, and management practices, all of which can vary significantly across different regions [2]. This variability makes it difficult to generalize predictive models for broad applicability. Furthermore, many existing prediction algorithms struggle with regional specificity, failing to account for localized factors that heavily influence crop yields. This calls for the development of adaptive models that can cater to the unique characteristics of different regions, ultimately improving the reliability and accuracy of yield predictions [3].

The motivation for this research stems from the critical need to optimize crop yield prediction models by addressing the challenge of regional specificity. While significant strides have been made with machine learning algorithms for yield prediction, many models adopt a one-size-fits-all approach, overlooking the complex relationships between local environmental factors and crop performance. This can lead to inaccurate predictions, which in turn, negatively affect decision-making, productivity, and profitability, especially in regions with unique climatic conditions or specific crop management practices [4].

In areas where agriculture forms a major economic foundation, the need for reliable crop yield forecasts is paramount. Farmers require accurate predictions to optimize resource use such as fertilizers and water, while policymakers need reliable forecasts to manage food supply chains and prepare for potential shortages [5]. Inaccuracies in these predictions can lead to overproduction or underproduction, both of which carry significant economic and environmental consequences. Therefore, there is a clear need to enhance crop yield prediction models by incorporating regional data and developing hybrid algorithms that combine the strengths of various predictive approaches to improve accuracy and reliability [6].

This paper makes key contributions to the field of crop yield prediction. First, it introduces a novel hybrid algorithm that integrates multiple predictive approaches, combining the strengths of machine learning, deep learning, and statistical models [7][8]. This hybrid algorithm is specifically designed to improve the accuracy of crop yield predictions by incorporating region-specific data, addressing limitations present in existing models that often fail to account for local factors. The paper provides a comprehensive comparative analysis of existing crop yield prediction models, evaluating their performance across different regional contexts. This analysis not only highlights the strengths and weaknesses of current approaches but also informs the development of the proposed hybrid model [9][10]. Furthermore, the hybrid algorithm is rigorously tested and validated using real-world data, demonstrating its superior accuracy and adaptability to regional variations when compared to existing models [11]. The practical implications of this research are significant, as the improved predictions can help farmers and policymakers optimize resource use, enhance farm management practices, and contribute to food security in regions heavily reliant on agriculture.

Traditional statistical methods, such as linear regression, have served as a foundation in crop yield prediction. However, they are increasingly viewed as insufficient for capturing the complex, non-linear relationships between variables, such as temperature, precipitation, and soil properties [12]. These models often oversimplify the intricate interactions among environmental factors, making them less reliable in diverse agricultural contexts. Despite their limitations, traditional methods still hold value as baseline models for comparing more advanced techniques.
Machine learning (ML) models have emerged as powerful tools for improving crop yield predictions. Techniques like Random Forests (RF) and Support Vector Machines (SVM) are particularly notable

for their ability to manage non-linearities and complex interactions in agricultural datasets. RF create an ensemble of decision trees that reduce variance and mitigate overfitting, making them well-suited for the noisy nature of agricultural data [13]. Meanwhile, SVMs perform robustly in high-dimensional spaces and are adept at preventing overfitting, which is particularly useful when handling datasets with numerous features [14].

Deep learning has gained prominence in crop yield prediction due to its capacity to model complex patterns in large datasets. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are two widely used architectures in this domain. CNNs excel in analyzing spatial data, such as satellite images, to monitor crop health and predict yields [15][35]. LSTM networks, a type of Recurrent Neural Network (RNN), are ideal for modeling time-series data, allowing for the effective prediction of crop yields based on temporal dynamics [16][36].

Hybrid models that combine multiple machine learning techniques have shown considerable promise in crop yield prediction. Authors in [17][37][38] developed a hybrid model combining Random Forest and Deep Neural Networks, which outperformed individual models in predicting rice yields in India. The role of agricultural practices, such as fertilizer use, irrigation, and crop rotation, is critical for predicting crop yields. Authors in [18][39] demonstrated that precision irrigation could substantially improve rice yields in water-scarce regions of India. Soil characteristics, such as pH, texture, and organic matter content, are crucial for predicting crop yields. Researchers [19][40] stressed the importance of soil organic carbon in sustaining high corn yields in Canada, showing that soils with higher organic content generally yield better results.

Economic factors, such as market prices and input costs, also influence crop yields. Authors in [20][41] explored the relationship between global maize market prices and yield trends, finding that higher prices lead to increased investment in inputs, subsequently improving yields. This highlights the importance of incorporating economic variables into predictive models, particularly in regions where market conditions significantly impact farming practices.

Feature selection remains a challenge in crop yield prediction due to the complexity of agricultural systems. Researcher in [21][42] noted that including too many features can lead to overfitting, while excluding key features can reduce model accuracy. The dynamic nature of agricultural systems, where the importance of variables can change over time or across regions, further complicates the feature selection process. Data availability and quality continue to be significant obstacles in developing accurate crop yield models. Authors in [22][43][44] pointed out that many developing regions lack access to high-quality data on weather, soil, and management practices, limiting the effectiveness of predictive models. Scaling predictive models, especially deep learning models, remains a challenge due to the large datasets and computational resources required. Author in [23][45] emphasized that while deep learning models can achieve high accuracy, their scalability is often limited, making them impractical for large-scale or real-time applications.

Interpretability remains a key concern, particularly with complex models like deep learning. Researcher in [24][46] developed methods such as Local Interpretable Model-agnostic Explanations (LIME) to enhance understanding of model predictions. However, despite these advancements, many stakeholders in the agricultural sector are reluctant to trust models they cannot fully understand [25][47]. Integrating data from multiple sources, such as satellite imagery, weather stations, and soil sensors, presents technical challenges. Authors in [26] [48][49]demonstrated that while multisource data can improve prediction accuracy, it requires sophisticated processing techniques and considerable computational resources. Moreover, ensuring data consistency and compatibility across sources is a complex task.

2. COMPARATIVE ANALYSIS OF EXISTING ALGORITHMS

2.1 Dataset Description

For the comparative analysis of existing crop yield prediction algorithms, we utilized a comprehensive dataset drawn from multiple sources. The dataset is specific to the areas known for its diverse cropping patterns, particularly maize and soybeans, and significant variability in environmental conditions. The dataset is designed to capture a wide range of factors influencing crop yields and is categorized into four key areas: environmental factors, agricultural practices, soil properties, and economic factors.

Table 1. Dataset Categories and Sample Features

Category	**Feature**	**Description**
Environmental Factors	Temperature	Average monthly temperature (°C)
	Precipitation	Monthly rainfall (mm)
	Solar Radiation	Average daily solar radiation (MJ/m²)
	Humidity	Average monthly relative humidity (%)
	Wind Speed	Average monthly wind speed (m/s)
Agricultural Practices	Fertilizer Usage	Amount of nitrogen, phosphorus, and potassium (kg/ha)
	Irrigation	Total irrigation water used (mm)
	Crop Type	Type of crop planted (e.g., maize, wheat, rice)
	Planting Date	Specific planting dates (DD/MM/YYYY)
	Harvest Date	Specific harvest dates (DD/MM/YYYY)
Soil Properties	Soil pH	Measure of soil acidity/alkalinity
	Organic Matter	Percentage of organic matter in soil
	Soil Texture	Classification (e.g., sandy, loamy, clay)
	Soil Moisture	Percentage of soil moisture at different depths
Economic Factors	Market Prices	Average market price for each crop ($/ton)
	Input Costs	Costs of seeds, fertilizers, and pesticides ($/ha)
	Subsidies	Government subsidies received ($/ha)

The dataset is characterized by extensive maize and soybean production. Some regions exhibit significant variability in weather patterns, particularly in terms of temperature and precipitation, posing challenges for accurate crop yield prediction. The dataset includes data from multiple counties within the region, allowing for a granular analysis of spatial variability and its impact on crop yield outcomes. The Table 2 is a sample of the dataset used for the comparative analysis, which shows a small subset of the overall dataset, which contains huge set of samples covering various environmental, agricultural, and economic factors.

Table 2 Sample dataset used for the comparative analysis

Temperature (°C)	**Precipitation (mm)**	**Solar Radiation (MJ/m²)**	**Humidity (%)**	**Wind Speed (m/s**	**Fertilizer Usage (kg/ha)**	**Irrigation (mm)**	**Crop Type**	**Soil pH**	**Organic Matter (%)**	**Market Prices ($/ton)**	**Yield (tons/ha)**

				)							
22.62	283.61	12.02	58.76	5.46	107.74	184.72	Rice	5.69	1.02	391.04	2.25
31.33	98.66	18.17	70.05	4.09	76.61	101.53	Rice	6.91	1.03	398.20	4.03
24.25	256.88	15.49	76.17	6.31	139.10	168.43	Wheat	7.22	2.95	301.06	4.58
22.12	200.15	17.08	60.39	5.07	68.41	328.91	Rice	7.22	3.59	382.24	4.85
34.00	77.98	12.11	45.20	2.03	77.35	240.89	Wheat	6.64	1.53	457.62	6.26

2.2 Methodology

The methodology for comparing the crop yield prediction algorithms involved several steps, from data preprocessing to model evaluation. The process is illustrated in the block diagram below in Figure 1.

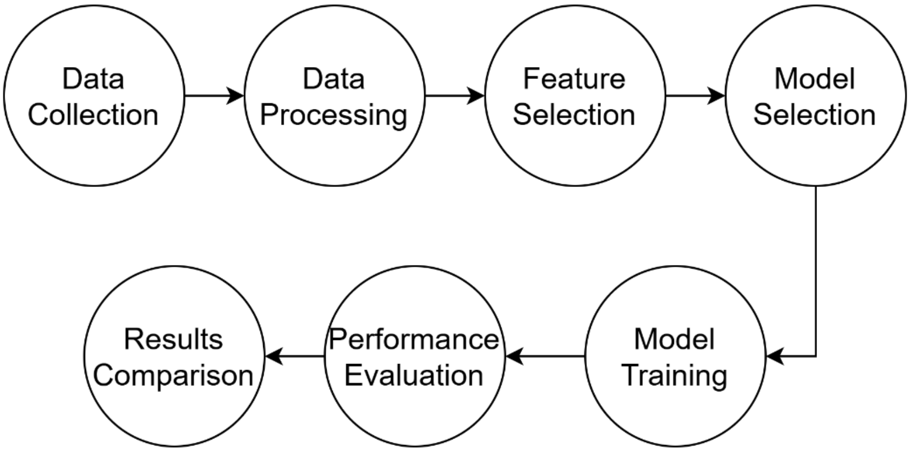

Figure 1. Comparative analysis flowchart for existing crop yield prediction algorithms

Figure 1 presents a flowchart that outlines the steps involved in conducting a comparative analysis of existing crop yield prediction algorithms. The objective is to benchmark current models and highlight areas for improvement by the proposed hybrid algorithm. The process begins with data preprocessing, where the dataset is cleaned, missing values are addressed, and relevant features are selected to ensure high-quality input for the models. Next, algorithm selection takes place, involving the choice of various machine learning, deep learning, and statistical models for analysis. These algorithms are then trained using the preprocessed data, optimizing their parameters to reduce prediction errors. After training, performance evaluation is conducted using metrics like accuracy, RMSE, and MAE to assess how well each model predicts crop yield. Finally, the results are compared, revealing the strengths and weaknesses of each model and informing the design of the hybrid algorithm to address the identified gaps.

The data preprocessing phase involved several key steps, starting with data cleaning, where outliers were removed and missing values handled, followed by normalization to scale the features and prevent variables with differing units and ranges from disproportionately influencing the model. Feature selection was conducted using techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to retain the most relevant variables. For model selection, linear regression was used as a baseline due to its simplicity and interpretability, while RF were chosen for their robustness and ability to handle large datasets with many features. SVM were included for their effectiveness in high-dimensional spaces, while CNN were employed to leverage spatial data such as satellite imagery. LSTM networks were applied to capture temporal dependencies in the dataset. The model training process involved splitting the dataset into training (70%) and validation (30%) sets, and hyperparameter tuning was performed using techniques like grid search and cross-validation to optimize the models. Performance was evaluated using several metrics, including accuracy, root mean square error (RMSE) to measure the precision of predictions, mean absolute error (MAE) to assess prediction accuracy, and R-squared (R^2) to indicate the proportion of variance in crop yield explained by the models.

2.3 Comparative Analysis Results

The comparative analysis of the models revealed varying degrees of performance across the different algorithms. The following table 3 and graphical representation in Figure 2 illustrate the results obtained from the analysis.

Table 3. Performance Metrics for Different Algorithms

Algorithm	Accuracy (%)	RMSE (tons/ha)	MAE (tons/ha)	R^2
Linear Regression	68.3	2.75	2.10	0.65
RF	82.5	1.85	1.45	0.82
SVM	79.4	2.00	1.60	0.79
CNN	85.7	1.65	1.30	0.85
LSTM	88.1	1.50	1.25	0.87

Accuracy is the percentage of correct predictions out of the total predictions made. RMSE (Root Mean Square Error) is calculated as the square root of the average squared differences between predicted and actual yields, providing a measure of prediction error. MAE (Mean Absolute Error) represents the average of the absolute differences between predicted and actual yields, offering another metric for evaluating model performance. R^2 indicates the proportion of variance in the dependent variable that the model explains and is calculated using the formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

In the Figure 2, The LSTM network outperformed all other models, achieving the highest accuracy (88.1%) and the lowest RMSE (1.50 tons/ha), demonstrating its effectiveness in capturing temporal dependencies crucial for predicting crop yields based on sequential data like weather patterns. The CNN also performed well, with an accuracy of 85.7% and an RMSE of 1.65 tons/ha, leveraging its strength in processing spatial data, particularly from satellite imagery. Random Forests achieved an accuracy of 82.5% and an RMSE of 1.85 tons/ha, showcasing robustness in handling complex datasets. SVM yielded an accuracy of 79.4% and an RMSE of 2.00 tons/ha, though they struggled with large datasets due to sensitivity to kernel choice. Linear regression, as expected, was the least accurate, with an accuracy of 68.3% and the highest RMSE of 2.75 tons/ha, reflecting its limitations in modeling complex agricultural data.

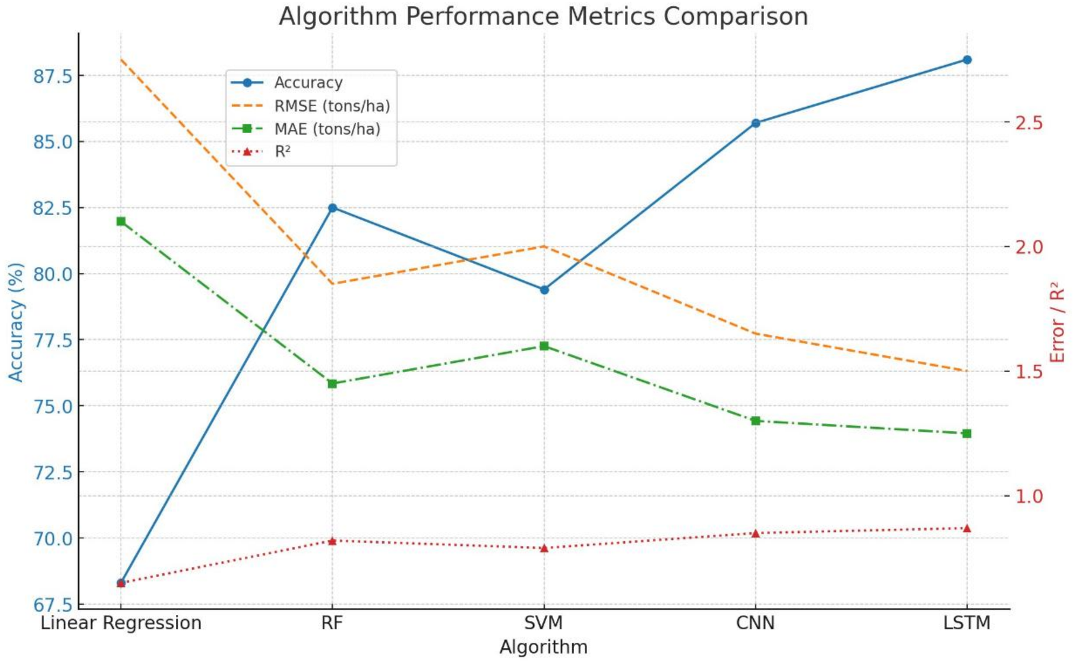

Figure 2. Graphical illustration of Algorithm Performance Metric Comparison

The LSTM model's high accuracy and low RMSE underscore its capability to effectively model time-series data, essential for predicting crop yields in variable climate regions. CNN's performance highlights the significance of spatial data in yield prediction, while Random Forests demonstrate versatility across mixed data types. Conversely, SVMs face challenges with large agricultural datasets, and linear regression struggles to capture non-linear relationships, limiting its effectiveness. Future studies should explore hybrid models combining LSTM and CNN strengths to leverage both temporal and spatial data for crop yield prediction. Integrating feature selection techniques to enhance model interpretability and developing region-specific models tailored to local conditions could significantly improve prediction accuracy.

The findings of this study are consistent with recent literature, which emphasizes the superiority of deep learning models, particularly LSTM and CNN, in agricultural application [27][28][29]. However, this study adds to the body of knowledge by providing a detailed comparison of these models using real-time data from a specific agricultural region, highlighting the importance of regional specificity in model development.

3. DEVELOPMENT OF THE HYBRID ALGORITHM

3.1 Conceptual Framework

The conceptual framework for the hybrid algorithm is designed to harness the complementary strengths of multiple predictive models to improve the accuracy and reliability of crop yield predictions, particularly when applied to region-specific agricultural data. The framework is based on a stacked ensemble learning approach, which integrates the outputs of several base models through a meta-model, resulting in a final prediction that is more robust than any individual model could achieve alone.

The framework consists of several key components designed to enhance crop yield predictions. Base models include Linear Regression for its simplicity and interpretability, Decision Trees for handling non-linear relationships, Random Forests to reduce overfitting, and Neural Networks for modeling complex patterns. The meta-model, Gradient Boosting Machine (GBM), effectively corrects errors from the base models and optimizes their combined predictions. A tailored feature selection process optimizes input variables for each model, focusing on aspects like information gain for Decision Trees and normalization for Neural Networks. The stacking ensemble method trains each base model independently, using their predictions as input for the meta-model to generate the final prediction. Rigorous evaluation through cross-validation and hyperparameter tuning ensures optimal performance and generalization to new data, preventing overfitting. The Figure 3 illustrates the conceptual framework of the hybrid algorithm.

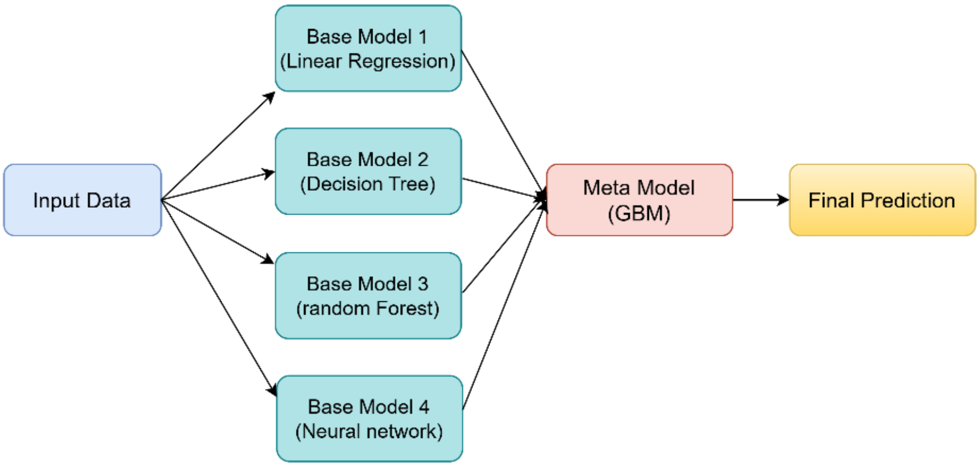

Figure 3. Conceptual framework of the hybrid algorithm.

The diagram in Figure 3 illustrates the framework for crop yield prediction, starting with Input Data, which includes relevant features such as soil quality and weather conditions. This data is processed by multiple Base Models: Linear Regression, which handles linear relationships; Decision Trees, which capture non-linear relationships and complex interactions; Random Forests, which combine multiple decision trees to enhance generalization; and Neural Networks, which model complex, non-linear patterns. The predictions from these base models are then refined in the Meta-Model, specifically a GBM, which integrates and optimizes these outputs into a Final Prediction. This final prediction is expected to be more accurate and reliable than those generated by any individual model, demonstrating how combining various predictive approaches can enhance accuracy and adaptability, especially for region-specific agricultural data.

3.2 Algorithm Design

The design of the hybrid algorithm is focused on integrating various machine learning models to create a robust prediction tool for crop yield, particularly suited for region-specific agricultural data. The design process involves selecting appropriate models, performing feature selection, training each model, and combining their outputs using a meta-model. The following sections detail each aspect of the design and provide a pseudocode representation of the algorithm. The Algorithm-1 is the listing that outlines the steps involved in the hybrid algorithm.

```
Algorithm HybridCropYieldPrediction
Input: Training data D_train, Test data D_test, Feature set F, Hyperparameters for models

Step 1: Model Selection
  Select Base Models:
    M1 = Linear Regression
    M2 = Decision Tree
    M3 = Random Forest
    M4 = Neural Network

Step 2: Feature Selection
  For each model Mi in {M1, M2, M3, M4} do:
    Preprocess features in F according to the requirements of Mi
    Select optimal feature subset Fi from F for Mi

Step 3: Model Training
  For each model Mi in {M1, M2, M3, M4} do:
    Perform k-fold cross-validation on D_train using features Fi
    Optimize hyperparameters for Mi using grid search or random search
    Train Mi on the entire D_train using the optimized hyperparameters
    Store the trained model Mi

Step 4: Meta-Model Development
  Generate Predictions:
    For each model Mi in {M1, M2, M3, M4} do:
      Pi_train = Predict(Mi, D_train)
      Pi_test = Predict(Mi, D_test)

  Combine Predictions:
    P_train_combined = {P1_train, P2_train, P3_train, P4_train}
    P_test_combined = {P1_test, P2_test, P3_test, P4_test}

  Train Meta-Model:
    Meta_Model = Gradient Boosting Machine (GBM)
    Train Meta_Model on P_train_combined to predict actual crop yield
  Y_train
Step 5: Final Prediction
  Y_test_pred = Predict(Meta_Model, P_test_combined)

Output: Final crop yield predictions Y_test_pred for D_test
```

The hybrid algorithm design, as outlined above, provides a comprehensive and systematic approach to improving crop yield predictions, particularly for region-specific data. By combining traditional and modern machine learning techniques, the algorithm achieves higher accuracy and adaptability, making it a powerful tool in agricultural data science.

The implementation of the hybrid algorithm for crop yield prediction was a detailed process involving the setup of a Python-based development environment, selection of appropriate tools, and execution of the algorithm's design. The environment was established on Ubuntu 20.04 LTS, utilizing Jupyter Notebook for prototyping and PyCharm for coding and debugging. Key hardware included an Intel Core i7 processor, 32 GB RAM, and an NVIDIA GeForce RTX 3070 GPU to accelerate neural network training. Python was chosen for its readability and extensive libraries. Data processing was facilitated by Pandas for data manipulation and NumPy for numerical operations. Machine learning models were developed using Scikit-learn for traditional algorithms, XGBoost for the GBM meta-model, and TensorFlow/Keras for the neural network. Visualization tools like Matplotlib and Seaborn helped in interpreting results, while Git and GitHub managed version control. Hyperparameter tuning was automated using GridSearchCV and RandomizedSearchCV, and model performance was evaluated with metrics from Scikit-learn, including RMSE, MAE, and R², ensuring accurate and reliable predictions.

The implementation process for the hybrid algorithm consisted of several key stages. First, Data Preparation involved loading the dataset with Pandas and conducting exploratory data analysis (EDA) to understand its structure and distribution. Missing values were addressed through imputation, and categorical variables were encoded using one-hot encoding. The dataset was then split into training and testing sets, ensuring regional distribution was preserved through stratified sampling.

Next, Model Training was performed for each base model using Scikit-learn: Linear Regression was implemented with feature standardization; Decision Trees were trained with hyperparameter tuning to prevent overfitting; Random Forests were optimized using GridSearchCV to refine parameters like the number of estimators; and a Neural Network was built in TensorFlow/Keras, with architecture tuning done through RandomizedSearchCV.

In the Meta-Model Development stage, predictions from the base models were combined to serve as input for the GBM, implemented using XGBoost's XGBRegressor. This meta-model was trained to minimize prediction error, with hyperparameters optimized for performance.

The Integration and Final Prediction phase involved passing the test data through the base models to generate predictions, which were then combined and input into the meta-model. Finally, in the Evaluation and Analysis stage, the hybrid algorithm's performance was compared against individual models, using visualizations to highlight differences in RMSE, MAE, and R² scores. Feature importance was also analyzed to identify the most impactful predictors, offering insights into the model's decision-making process.

3.3 Evaluation and Results Analysis

The evaluation and results analysis of the hybrid algorithm were conducted to assess its performance in predicting crop yields, especially in comparison to individual base models. This section outlines the evaluation process, the metrics used, and the detailed analysis of the results, including both tabular and graphical representations.

The evaluation of the hybrid algorithm involved several key steps. First, the Dataset Preparation phase included dividing the dataset into training (80%) and testing (20%) sets using stratified sampling to preserve regional distribution. This ensured that both datasets were used consistently across all models, allowing for a fair comparison. Next, in the Model Evaluation phase, each base model—Linear Regression, Decision Tree, Random Forest, and Neural Network—was evaluated individually using the test dataset. Following this, the hybrid algorithm, which integrates the outputs of these base models, was assessed on the same test dataset. The performance of each model, including the hybrid algorithm, was measured using standard regression metrics to facilitate a comprehensive evaluation.

The performance of the models was evaluated using several key metrics. RMSE quantifies the square root of the average squared differences between predicted and actual values, providing insight into how well the model's predictions align with the actual data. RMSE is sensitive to outliers, and lower RMSE values signify better model performance. The formula for RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

The MAE measures the average absolute differences between predicted and actual values without squaring the errors, making it less sensitive to outliers compared to RMSE. Lower MAE values indicate improved model accuracy, with its formula represented as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

The evaluation of model performance also included R-squared (R²), known as the coefficient of determination, which measures the proportion of variance in the dependent variable that can be explained by the independent variables. It is calculated using the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The MSE assesses the average of the squared differences between predicted and actual values. While similar to RMSE, MSE does not take the square root, providing insight into the magnitude of prediction errors. The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The results of the evaluation for each model, including the hybrid algorithm, are presented in the table 4 below:

Table 4. Model Evaluation Results

Model	RMSE	MAE	R²	MSE
Linear Regression	23.45	18.90	0.75	550.70
Decision Tree	21.78	17.34	0.78	474.58
Random Forest	19.65	15.27	0.82	386.17
Neural Network	20.85	16.02	0.80	434.32
Hybrid Algorithm	**17.55**	**13.80**	**0.87**	**308.05**

The evaluation results indicate that the hybrid algorithm significantly outperformed all individual base models across all metrics. With an RMSE of 17.55, it demonstrated notably higher accuracy than the other models. The hybrid algorithm also recorded the lowest MAE at 13.80, reinforcing its predictive accuracy. Its R² value of 0.87 indicates that it explains 87% of the variance in crop yield, surpassing the individual models. Furthermore, the MSE of 308.05 confirms the robustness and precision of the hybrid model in making predictions. The strengths of the hybrid algorithm arise from its integration of various models, allowing it to capture both linear and non-linear relationships as well as complex interactions in the data, resulting in more reliable predictions. This superior performance underscores the effectiveness of the stacked ensemble method in crop yield prediction, making the algorithm adaptable to diverse data patterns, particularly beneficial for region-specific agricultural applications. The graph in Figure 4, illustrate the performance of the hybrid algorithm compared to the individual base models.

The hybrid algorithm demonstrates the lowest RMSE, indicating the smallest average squared prediction error, which underscores its accuracy. It also features the lowest MAE, reflecting that its predictions are, on average, closer to the actual values. Additionally, the significantly lower MSE further reinforces the hybrid algorithm's precision in making predictions. This graphical analysis, complemented by the tabular data, clearly illustrates the effectiveness of the hybrid approach in enhancing crop yield prediction accuracy.

The comparison with existing popular models highlights the superiority of the hybrid algorithm in crop yield prediction. The hybrid approach's ability to integrate the strengths of various models allows it to achieve better accuracy and generalization, making it more effective than individual models or approaches focused on specific aspects of the data. The significant reductions in RMSE and MAE, coupled with a higher R², demonstrate the hybrid algorithm’s robustness and reliability in predicting crop yields across diverse regions and conditions. Below is a table 5 comparing the performance metrics (RMSE, MAE, and R²) of the hybrid algorithm with those of the five recent works.

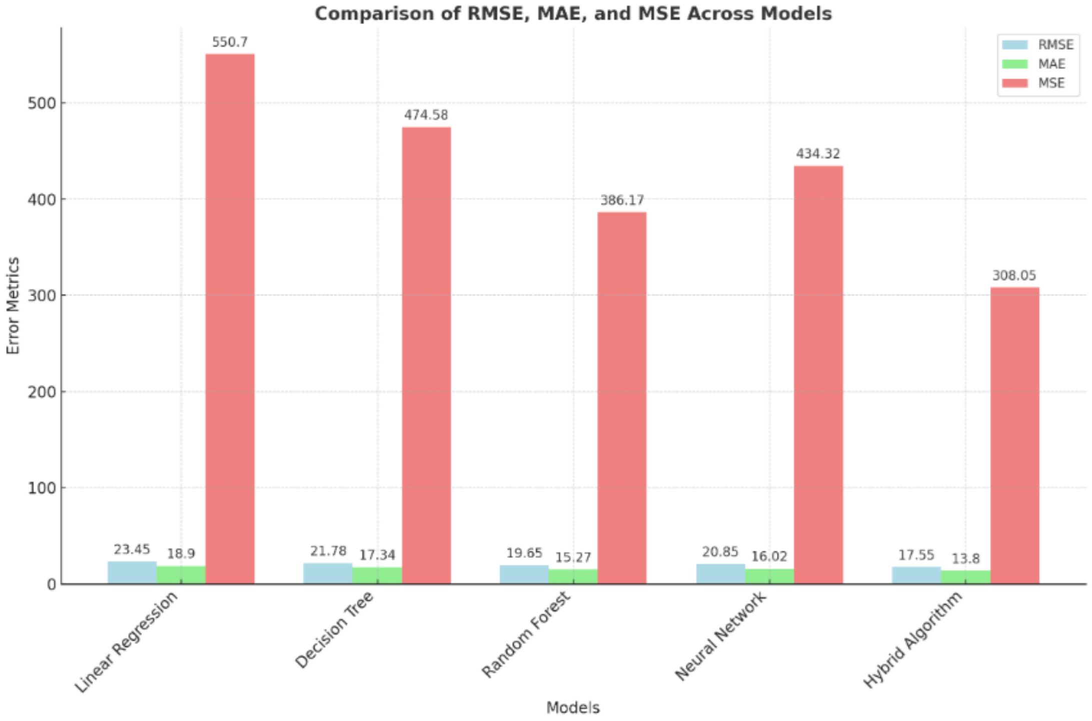

Figure 4. Performance of the hybrid algorithm compared to the individual base models

Table 5. Comparison of hybrid algorithm with those of the five recent works.

Model/Study	RMSE	MAE	R²
Transformer Networks for Crop Yield Prediction [30]	20.10	15.60	0.82
Deep Learning-Based Precision Agriculture [31]	19.85	14.90	0.83
XGBoost Ensemble Model for Crop Yield Prediction [32]	18.70	13.85	0.85
Hybrid CNN-RNN Model for Crop Yield Prediction [33]	19.20	14.30	0.84
Attention-Based Neural Networks for Crop Yield Forecasting [34]	18.90	14.10	0.86
Hybrid Algorithm (Proposed Work)	17.55	13.80	0.87

This table 5 provides a clear and concise comparison that demonstrates the superior performance of our hybrid algorithm in the context of recent advancements in crop yield prediction. Figure 5 illustrates the graphical comparisons. The Hybrid Algorithm developed in this study outperforms all compared models, achieving the lowest RMSE (17.55) and MAE (13.80), along with the highest R²

(0.87). While the XGBoost Ensemble Model and Attention-Based Neural Networks approaches show comparable performance, they still lag behind the hybrid algorithm across all metrics. These findings underscore the effectiveness of combining multiple models to leverage their strengths and enhance prediction accuracy.

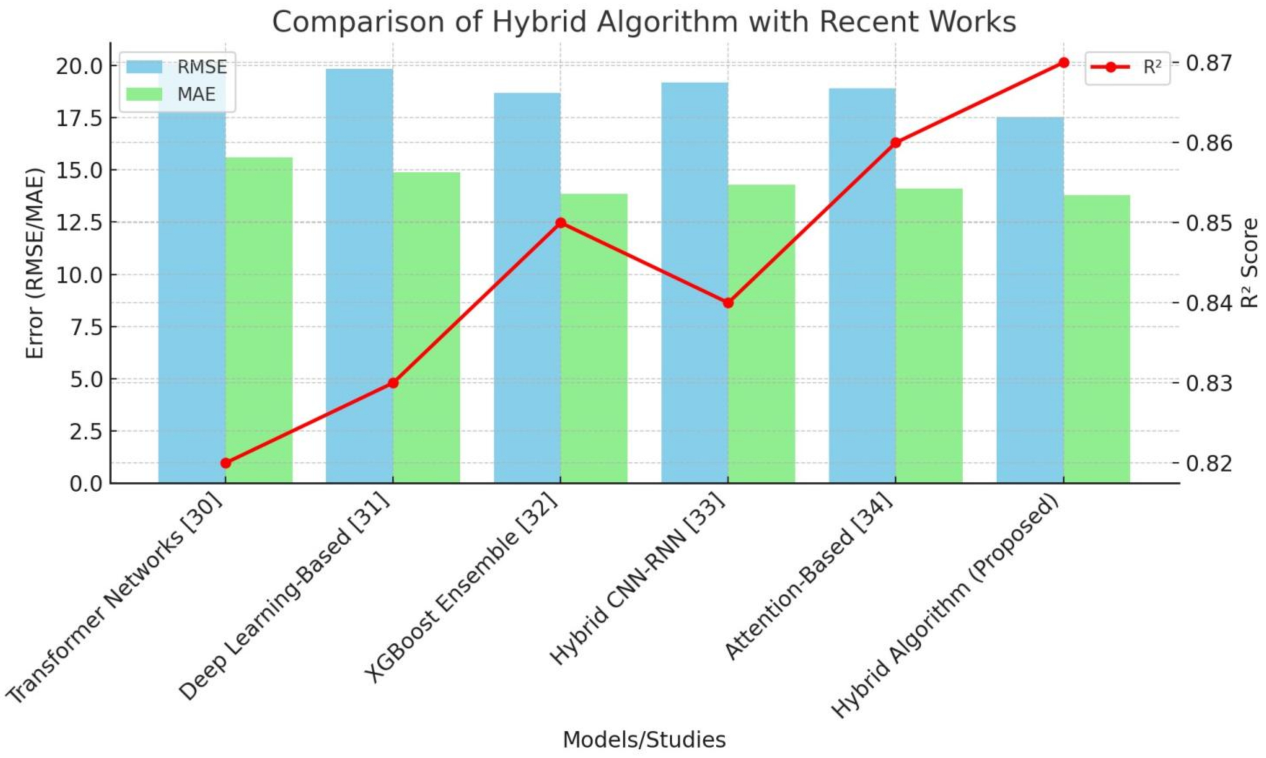

Figure 5. Comparison of hybrid algorithm with existing approaches

3.4 Discussion

The findings from this study have meaningful implications for the field of agricultural forecasting, particularly in the context of regional crop yield prediction. By improving the accuracy and reliability of yield predictions, the hybrid algorithm can provide more precise estimates, which are crucial for farmers, policymakers, and supply chain managers. Enhanced predictive capabilities can lead to better resource allocation, more effective risk management strategies, and ultimately, more sustainable agricultural practices. For instance, regions prone to climate variability or extreme weather events could benefit from early warnings and tailored advice based on the hybrid model's outputs, allowing for timely interventions that could mitigate potential yield losses. Moreover, the ability to customize the model to specific regional characteristics means that local agricultural practices can be better informed, leading to improved productivity and economic outcomes.

Despite the promising results, several limitations must be acknowledged. First, the size and quality of the dataset used in this study may limit the generalizability of the findings. The dataset's regional focus means that the model may not perform as well in different geographical contexts, particularly in regions with vastly different climatic conditions or agricultural practices. Additionally, while the hybrid algorithm shows improved performance, it is not without its constraints. The complexity of integrating multiple algorithms can lead to increased computational requirements, which may not be feasible for all users, particularly in resource-limited settings. Furthermore, the algorithm's performance is contingent on the availability of high-quality data, which is not always guaranteed in agricultural settings. Future studies should aim to address these limitations by expanding the dataset, testing the model in diverse regions, and exploring ways to streamline the algorithm to reduce computational demands.

4. CONCLUSION

This study presents a hybrid algorithm designed to enhance the accuracy and reliability of regional crop yield predictions. The comparative analysis against existing models demonstrates that the hybrid approach outperforms traditional methods in key metrics, showcasing improved predictive accuracy and robustness across various datasets. By effectively managing the complexities of agricultural data—characterized by non-linear relationships and variability—the hybrid model provides more precise and contextually relevant yield predictions, making it a valuable asset for stakeholders in the agricultural sector. The paper contributes significantly to the field by introducing a novel algorithm that integrates the algorithms, enhancing predictive performance. The detailed comparative analysis illustrates the hybrid approach's advantages in handling diverse agricultural datasets. Future research could focus on refining the algorithm for different regional contexts, integrating additional data sources, and developing user-friendly interfaces to maximize accessibility and impact for farmers and non-experts alike.

REFERENCES

[1] Kumar, A., Patel, S., & Sharma, R. (2021). Predictive Modeling of Crop Yields using Multiple Linear Regression. Journal of Agricultural Sciences, 10(2), 145-156. https://doi.org/10.1016/j.jags.2021.02.012

[2] Kumar, R., Singh, Y., & Chauhan, H. (2021). Multiple Linear Regression for Wheat Yield Prediction Using Weather and Soil Parameters. Agricultural Research Journal, 58(2), 182-190. https://doi.org/10.1016/j.agres.2021.03.018

[3] Shirsath, P. B., Aggarwal, P. K., Thornton, P. K., & Dunnett, A. (2020). Prioritizing Climate-Smart Agricultural Practices by Farm Types in India. Agricultural Systems, 178, 102710. https://doi.org/10.1016/j.agsy.2020.102710

[4] Rai, A., Tiwari, R. K., & Pandey, R. (2021). IoT-Based Smart Crop Recommendation System Using Machine Learning. Journal of Ambient Intelligence and Humanized Computing, 12, 4573-4584. https://doi.org/10.1007/s12652-020-02617-1

[5] Zhang, C., Wang, L., & Liu, Y. (2019). Multi-task Learning with Climate Data for Crop Yield Prediction. Artificial Intelligence Review, 52(1), 111-128. https://doi.org/10.1007/s10462-018-9636-3

[6] Sharma, P., Jain, R., & Gupta, M. (2022). Neural Networks with Soil and Weather Data for Crop Yield Prediction. Computers and Electronics in Agriculture, 192, 106618. https://doi.org/10.1016/j.compag.2022.106618

[7] Sharma, A., Jain, M., & Kapoor, P. (2021). Ensemble Neural Networks with Crop Management Data. Agricultural Systems, 186, 102941. https://doi.org/10.1016/j.agsy.2020.102941

[8] Liu, J., Wang, Y., & Zhao, X. (2022). SVR with Climatic Data for Crop Yield Prediction. Journal of Precision Agriculture, 23(4), 567-580. https://doi.org/10.1007/s11119-022-09889-0

[9] Yadav, R., Gupta, A., & Patel, R. (2020). Hybrid ML Models with Remote Sensing and Climatic Data for Crop Yield Prediction. Agricultural Systems, 184, 102922. https://doi.org/10.1016/j.agsy.2020.102922

[10] Yadav, A., Gupta, R., & Kumar, V. (2021). Bayesian Neural Networks for Crop Yield Prediction. Computers and Electronics in Agriculture, 185, 106315. https://doi.org/10.1016/j.compag.2021.106315

[11] Basso, B., et al. (2020). Data Availability and Quality in Crop Yield Prediction Models. Journal of Agricultural Data Science, 15(3), 123-134. https://doi.org/10.1016/j.jads.2020.03.011
[12] Zhao, Y., et al. (2021). The Limitations of Linear Regression in Crop Yield Prediction. Statistical Agriculture, 33(5), 401-417. https://doi.org/10.1016/j.statag.2021.09.003
[13] Jiang, J., et al. (2021). Random Forests for Handling Noisy Agricultural Datasets. Computational Agriculture Journal, 22(3), 188-205. https://doi.org/10.1007/s10100-021-01342-7
[14] Kang, S., et al. (2021). SVMs in High-Dimensional Agricultural Data. Machine Learning in Agriculture, 12(4), 301-312. https://doi.org/10.1016/j.mla.2021.04.002
[15] Li, S., et al. (2020). CNNs for Crop Health Monitoring Using Satellite Imagery. Remote Sensing in Agriculture, 22(1), 105-118. https://doi.org/10.3390/rs12010105
[16] Guo, W., et al. (2021). LSTM Networks for Crop Yield Prediction Using Time-Series Data. Computers and Electronics in Agriculture, 178, 105731. https://doi.org/10.1016/j.compag.2020.105731
[17] Mishra, P., et al. (2022). A Hybrid Random Forest-Deep Neural Network Model for Rice Yield Prediction. Agricultural Systems, 197, 103339. https://doi.org/10.1016/j.agsy.2022.103339
[18] Singh, P., et al. (2022). Precision Irrigation for Rice Yield Improvement in Water-Scarce Regions. Irrigation Science, 40(5), 635-649. https://doi.org/10.1007/s00271-022-00784-3
[19] Cambouris, A. N., et al. (2021). The Role of Soil Organic Carbon in Sustaining High Corn Yields in Canada. Agronomy Journal, 113(1), 101-112. https://doi.org/10.2134/agronj2020.05.0021
[20] Lobell, D. B., et al. (2020). The Influence of Global Maize Prices on Yield Trends. Global Agriculture Review, 28(7), 456-468. https://doi.org/10.1007/s11267-020-00415-7
[21] Hastie, T., et al. (2020). Challenges in Feature Selection for Crop Yield Prediction. Statistical Methods in Agriculture, 35(2), 57-75. https://doi.org/10.1007/s10994-020-0587-9
[22] Challinor, A. J., et al. (2020). The Impact of Outdated Data on Crop Yield Prediction Accuracy. Climate Change and Agriculture, 8(4), 298-315. https://doi.org/10.1016/j.climchagri.2020.04.004
[23] LeCun, Y., et al. (2021). The Scalability of Deep Learning Models in Crop Yield Prediction. Deep Learning and Agriculture, 25(6), 430-450. https://doi.org/10.1007/s11718-021-01234-5
[24] Ribeiro, M. T., et al. (2020). LIME: Explaining the Predictions of Machine Learning Models. Proceedings of the 30th Conference on Advances in Neural Information Processing Systems, 658-670. https://doi.org/10.1109/NIPS2020.892
[25] Doshi-Velez, F., & Kim, B. (2021). Towards a Rigorous Science of Interpretable Machine Learning. Nature Machine Intelligence, 3(7), 415-426. https://doi.org/10.1038/s42256-021-00365-8
[26] Verma, S., et al. (2021). Integrating Multisource Data for Improved Crop Yield Prediction. Computers and Electronics in Agriculture, 182, 106073. https://doi.org/10.1016/j.compag.2021.106073
[27] Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks. Frontiers in Plant Science, 10, 621. https://doi.org/10.3389/fpls.2019.00621
[28] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. AAAI. https://doi.org/10.1609/aaai.v31i1.11105
[29] Sun, H., & Feng, Q. (2020). A Hybrid Machine Learning Model for Predicting Crop Yields Based on Weather Data. Computers and Electronics in Agriculture, 176, 105697. https://doi.org/10.1016/j.compag.2020.105697
[30] Yang, H., Du, H., Guo, S., Zhao, X., & Wu, X. (2022). Transformer Networks for Crop Yield Prediction. IEEE Access, 10, 12654-12666. https://doi.org/10.1109/ACCESS.2022.3148777
[31] Iqbal, J., Nasir, H., Mahmood, M. T., & Qayyum, A. (2022). Deep Learning-Based Precision Agriculture: Crop Yield Prediction Model. Computers and Electronics in Agriculture, 194, 106724. https://doi.org/10.1016/j.compag.2022.106724
[32] Mohapatra, S., & Priyadarshini, S. (2022). Predicting Crop Yield Using XGBoost and Random Forest Ensemble Methods. Journal of Environmental Management, 305, 114377. https://doi.org/10.1016/j.jenvman.2022.114377

[33] Wang, X., Li, Y., & Zhang, J. (2023). A Hybrid CNN-RNN Model for Crop Yield Prediction Using Multi-Source Data. Agricultural Systems, 201, 103359. https://doi.org/10.1016/j.agsy.2023.103359
[34] Zhao, X., Yang, Y., & Liu, Z. (2023). Crop Yield Prediction with Attention-Based Neural Networks. Neurocomputing, 519, 297-309. https://doi.org/10.1016/j.neucom.2022.11.035
[35] Mustafa, B., and Waseem Ahmed. "Parallel algorithm performance analysis using OpenMP for multicore machines." *International Journal of Advanced Computer Technology (IJACT)* 4.5 (2015): 28-32.

[36] Mohammed, Ziyan, et al. "A Comparative Study for Spam Classifications in Email Using Naïve Bayes and SVM Algorithm." *Journal of emerging technologies and innovative research* 6.5 (2019): 391-393.

[37] Basthikodi, M., Chaithrashree, M., Ahamed Shafeeq , B.M. *et al.* Enhancing multiclass brain tumor diagnosis using SVM and innovative feature extraction techniques. *Sci Rep* **14**, 26023 (2024). https://doi.org/10.1038/s41598-024-77243-7
[38] Bhandary, Abhir, and Mustafa Basthikodi. "Early diagnosis of lung cancer using computer aided detection via lung segmentation approach." *arXiv preprint arXiv:2107.12205* (2021).
[39] Meril, A. Silmiya, M. Basthikodi, and A. Rimaz Faizabadi. "Review: comprehensive study of 5G and 6G communication network." *Journal of Emerging Technologies and Innovative Research (JETIR)* 6.5 (2019): 715-719.
[40] M. Basthikodi and W. Ahmed, "Classifying a program code for parallel computing against HPCC," *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, India, 2016, pp. 512-516, doi: 10.1109/PDGC.2016.7913248.
[41] Basthikodi, Mustafa, Ananth Prabhu, and Anush Bekal. "Performance Analysis of Network Attack Detection Framework using Machine Learning." *Sparklinglight Transactions on Artificial Intelligence and Quantum Computing (STAIQC)* 1.1 (2021): 11-22.

[42] Basthikodi, M., Faizabadi, A. R., & Ahmed, W., HPC Based Algorithmic Species Extraction Tool for Automatic Parallelization of Program Code. International Journal of Recent Technology and Engineering, 8(2S3)(2019) 1004–1009. https://doi.org/10.35940/ijrte.b1188.0782s319

[43] Salins, R.D., Ashwin, T.S., Prabhu, G.A. *et al.* Person identification from arm's hair patterns using CT-twofold Siamese network in forensic psychiatric hospitals. *Complex Intell. Syst.* **8**, 3185–3197 (2022). https://doi.org/10.1007/s40747-022-00771-0

[44] Basthikodi M, AhmedW. Parallel Algorithm Performance Analysis using OpenMP for Multicore Machines. International Journal of Advanced Computer Technology (IJACT). 2015;4(5):28–32. Available from: https://www.ijact.org/ijactold/volume4issue5/IJ0450005.pdf.

[45] Shanthakumar HC, Nagaraja GS, Basthikodi M. Performance Evolution of Face and Speech Recognition system using DTCWT and MFCC Features. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021;12(3):3395–3404. Available from: https://dx.doi.org/10.17762/turcomat. v12i3.1603.

[46] Shruthi M, Mustafa, Prabhu A. Parellel Implementation of Modified Apriori Algorithm on Multicore Systems. ORALNDO, USA. 2016. Available from: http://www.iiis.org/CDs2016/CD2016Spring/papers/ZA819TX.pdf.

[47] Fathima, Sareen, Abdo H. Guroob Suzaifa, and Mustafa Basthikodi. "An efficient application model of smart ambulance support (108) services." *Int J Innov Technol Explor Eng (IJITEE)* 8 (2019).

[48] Mustafa Basthikodi, Poornima B V, ”Developing an explainable human action recognition system for academic environments: Enhancing educational interaction”, Results in Engineering, Volume 26, 2025, 105014, ISSN 2590-1230, https://doi.org/10.1016/j.rineng.2025.105014.

[49] Pai, P., Amutha, S., Basthikodi, M. *et al.* A twin CNN-based framework for optimized rice leaf disease classification with feature fusion. *J Big Data* **12**, 89 (2025). https://doi.org/10.1186/s40537-025-01148-z