

Machine Learning Based Method for Forecasting Crop Yield

Renu Kumari¹ , Vikash Sawan² , Mrinmoy Kayal³ 

¹Research Scholar K. K. University, Nalanda Bihar

^{2,3}Assistant Professor, Department of Computer Engineering & Applications GLA University, Mathura Uttar Pradesh.

*Corresponding Email: vikash.sawan@gla.ac.in



This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Applications of machine learning are revolutionizing data processing and decision-making, which is having a significant effect on the global economy. Given the worldwide food supply crisis, agriculture is one of the industries where the effects are most noticeable. This paper focuses on crop yield prediction based on pattern analysis with the help of the machine learning approach, which focuses on data acquisition, preprocessing, and assessment. Taking the Crop Yield Prediction Dataset as a solution, the most potential factors, including rainfall, temperature, and pesticide, have been identified to have the most influential factors in creating better prediction models. Among these, Decision Trees, Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes, and Long Short-Term Memory (LSTM) are the most common and are checked for their effectiveness. It brings out facts that are instrumental in analysis to improve the yields on farms and come up with possible recommendations on precision farming and sustainable agriculture. This paper aims to provide insights that can help improve farm yields.

Keywords: Crop yield prediction, Machine learning, Precision agriculture, Sustainable farming.

1. Introduction

Agriculture plays a vital role in ensuring food security, economic support, and environmental sustainability. With increasing population across the globe, demand for food production increases and it becomes important to use advanced technologies to enhance agriculture's productivity. Accurate predictions of crop yields have been agriculture's biggest challenge for farmers and agriculture specialists because there are numerous factors such as soil health, weather patterns, and farming practices affecting production. Historical records, weather forecasts, and statistical modeling have been traditionally used to predict crop yields [1]. These do not provide correct predictions because agriculture has a dynamic nature and complexity involved. The use of machine learning has emerged as a breakthrough to support decision-making with a focus on enhancing crop yield predictions. Through processing huge amounts of data, ML algorithms can identify patterns, analyse environmental factors, and provide better predictions than conventional methods [3],[18],[21]. This paper explains how machine learning has been utilized in crop yield prediction using different ML methods, techniques to obtain data, and preprocessing techniques. It also discusses challenges, practical applications,

ethics, and future directions in precision agriculture. The objective is to provide a holistic overview of how ML has changed farm practices in recent years, increasing productivity, reducing wastage, and making agriculture sustainable [4].

Advances in machine learning and crop simulation modeling have opened up new avenues for agricultural prediction. These technologies have each brought distinct capabilities and considerable gains in prediction performance; however, they have been primarily evaluated independently, and there may be benefits in integrating them to further increase prediction accuracy [7].

2. Literature Review

2.1 Historical Yield Prediction Techniques

According to Kallenberg et al., 2023[6], Previous crop production forecast techniques were predominantly multiple linear regression models, time series analysis, and agro-meteorology-based models. These methods utilized earlier records in soil fertility, climate, and cultivation to make production forecasts. These methods could not account for fluctuations in climate, pests, and soil and were therefore not very reliable and accurate. Traditional methods required a lot of manually collected data and were time and labour-consuming.

2.2 Deficiencies in Conventional Methods

According to Bassine et al., 2023[2], one of the largest shortcomings with traditional predictive techniques is that they can't effectively deal with high-dimensional, complex data. With more and more data coming from IoT sensors, satellite imagery, and real-time monitoring of climate, traditional models can't effectively deal with these different sources. Moreover, statistical models rely on linear relationships among factors, whereas real agronomic factors have nonlinear interdependencies, making traditional predictions ineffective.

2.3 Development in Machine Learning in Agriculture

According to Shook et al., 2021[10], Machine learning has revolutionized crop yield predictions by enabling adaptive, automatic, and highly effective predictive models. Unlike conventional methods, ML algorithms can handle enormous amounts of information, learn from them, and improve predictive performance with time. Supervised algorithms such as Decision Trees, Support Vector Machines (SVM), and Random Forests have been widely applied in agriculture forecasting.

Deep learning algorithms such as Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) are more effective in processing satellite imagery and remote sensing. One of the strongest arguments in favor of ML-based methods lies in their ability to accept real-time inputs from IoT-connected farms and consolidate different inputs such as soil moisture levels, temperature fluctuations, and crop growth rates [9].

2.4 Recent developments in Machine Learning for Crop Yield Prediction

According to Zhou et al., 2022[12], Machine learning has considerably improved crop yield forecasting effectiveness and accuracy. Methods integrating remote sensing, geospatial analysis, and climate modelling

have been investigated to enhance predictive modelling. Deep learning methods with huge amounts of data analysis have enabled real-time forecasting, and farmers can now react in real-time with real-time environmental inputs. Hybrid machine learning algorithms such as Convolutional Neural Networks and Support Vector Machines have been developed to enhance predictive performance. Explainable AI (XAI) has also been utilized in precision agriculture to establish credibility in AI decision-making. Cloud, NLP edge computing and AI [19],[20] have also enabled more effective processing with high amounts of data.

3. Machine Learning Techniques for Crop Yield Estimation

Machine learning has proved to be a major facilitator in agriculture forecasting, particularly in crop yield forecasting with high efficiency and accuracy. Unlike traditional methods, ML algorithms can analyse huge volumes of data, identify hidden patterns, and give predictions from real-time inputs.

3.1 Supervised Learning Techniques

Supervised learning involves utilizing labeled datasets to train a model to enable it to make predictions from historic agriculture information.

- Decision Trees (DTs): DTs partition variables in agriculture by creating a tree-like model in which decisions are made based on feature importance, i.e., soil nutrients, weather patterns, and crop type (Shah Hosseini et al., 2021) [8]. They offer interpretability and transparency and can be susceptible to overfitting in high-complexity data.
- Random Forest: A technique in ensemble learning that builds a collection of DTs and combines their predictions to improve accuracy. RF models handle noisy agronomic data efficiently and can effectively work with massive environmental datasets.
- Support Vector Machines (SVMs): SVMs are applied to provide classification and regression analysis by selecting the best hyperplane to separate different crop yield outputs. They perform well with small sample sets but can be computationally demanding in high-sample applications.

3.2 Deep Learning Techniques

Deep learning techniques have revolutionized crop yield predictions by employing neural networks to examine high-dimensional inputs like satellite imagery, sensor readings, and predicted weather. Some such prominent deep-learning techniques include:

- Artificial Neural Networks (ANNs): ANNs mimic brain processes and have widespread applications in predicting yield by learning complex, nonlinear relationships between factors such as soil, climatic factors, and cultivation practices.
- Convolutional Neural Networks (CNNs): CNNs emphasize spatial analysis and can be used in satellite imagery and remote sensing to study crop growth and disease detection (Wang et al., 2024) [11].
- Recurrent Neural Networks: One variety of Recurrent Neural Networks, LSTMs work very effectively in handling long time-series data. Therefore, they prove to be very effective in crop production predictions in different cultivation seasons.

3.3 Hybrid Approaches

Hybrid models use a combination of various ML algorithms to enhance predictive power by combining each method's strongest points. Some examples include:

- Merging RF with ANNs: This approach combines decision trees' interpretability with neural networks' flexibility to improve predictive performance.
- Integration of CNNs and SVMs: CNNs extract spatial features from satellite imagery and SVMs classify patterns in extracted information. This integration improves remote sensing-based yield estimation.
- IoT-ML Integration: Utilization of Internet of Things (IoT) sensors to obtain real-time data about soil health, temperature, moisture, and plant growth has improved predictions using ML. Sensor-based inputs and ML algorithms can be blended to give predictive models real-time estimates.

4.0 Data Collection and Preprocessing for Yield Prediction

Data collection and preprocessing are critical processes in building strong machine learning (ML) models for crop yield forecasting. The performance and dependability of ML algorithms depend to a considerable extent on the quality, diversity, and quantity of training data. Precise preprocessing of data ensures that ML algorithms can extract useful patterns from raw farm-level data, leading to better predictions and decision-making.

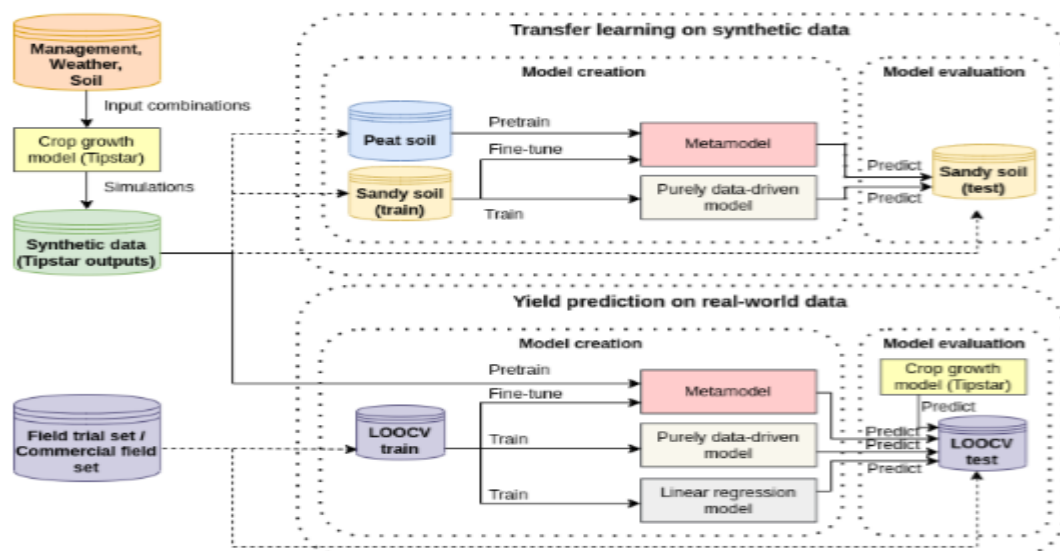


Figure 1: Data flow and model development

4.1 Crop Yield Prediction Data Sources

For this research, the Crop Yield Prediction Dataset was considered and adopted as the major dataset. There are 28,242 records from which seven vital fields compile several agricultural and environmental aspects

(JASWAL, 2024) [5]. The following is a narration of the most important features of the data set that has been explained in brief in the preceding sections of this paper:

- **Area (Country):** This column consists of nominal data as it gives distinction to other various countries or regions. Because the conditions of agriculture differ from one area to another, this variable is deemed very essential when analysing the effects of the environment on the production of crops.
- **Item:** A list of various agricultural products that are offered in this company: Crops also have different growth progresses, demands for fertilizers, and climate sensitivity, which the model must discern or recognize.
- **Year :** An integer field that contains the year value for the record entered. This helps to track them over a certain period, and the ML model incorporates seasonal or elapsed time changes in farming.
- **hg/ha_yield:** This column portrays the crop yield in the form of hectograms per hectare (hg/ha), which is always an important measure of efficiency in farming. This is the variable which is of focus for the prediction task in the ML model.
- **average_rain_fall_mm_per_year:** The type of allowance is a float column representing the average annual rainfall measurement in millimetres for the countries of interest. Rainfall usually has a significant impact on the yield and thus is an important feature in the model.
- **pesticides_tonnes:** A float column that contains the number of pesticides that have been used in tonnes. This feature would be useful when evaluating a model since excessive use of pesticides increases yields but comes with environmental risks.
- **avg_temp :** This column contains float data as it shows mean temperature in degrees Celsius of each country in the respective year. This paper specifically focuses on analysis of the effects of temperate on crop growth and yield prediction since temperature affects crop growth.

4.2 Data Pre-processing Techniques

The data was pre-processed by undertaking some data cleaning exercises to enhance the quality of the data for modelling. The steps included:

- **Data Cleaning:** To ensure the data is clean, procedures such as removing duplicate data and dealing with the missing data were made. They also experienced some missing values in some of the rainfall, temperature, and pesticide data and hence have been replaced by mean imputation to reduce the missing values.
- **Feature Engineering and Selection:** These included such values as rainfall, temperature and the usage of pesticide, while data that may not have been important or repetitive were removed. The engineered features highlighted climatic patterns and farming techniques that are positively associated with the production of crops.
- **Data Scaling and Normalization:** This feature involved such as rainfall, temperature, and yield vary in different scales, data normalization techniques were used to normalize the data. This encouraged none of them to dominate the others in making contributions during the model training process.

- **Data Augmentation:** Some regions and some crop types have limited data, especially in the past period, so several techniques for synthetic data were used to fill the existing gaps. This boosted the model accuracy by enhancing the number of records in areas that don't contribute many records in the fields of agriculture.

4.3 Importance of Pre-processing Data

Data pre-processing enhances ML model robustness, efficiency, and interpretability. If not addressed, ML algorithms may produce biased, inaccurate, or uninformative predictions, and these may limit their application in real-world agriculture.

5.0 Results & Performance Evaluation

To understand how well different machine learning algorithms can perform in predicting crop yields, a set of classification algorithms was compared against the agriculture datasets.

- Accuracy: It measures how correct predictions generally are.
- Precision: It describes how many predicted positives were correct.
- Recall refers to whether a model can cover all possible examples.
- F1 score: Harmonic mean between recall and precision, balancing both.

5.1 Performance Metrics of Different ML Models

Table 1: ML Algorithms Performance Metrics

| Model | Accuracy (%) | Precision | Recall | F1 Score |
|---------------------------------|--------------|-----------|--------|----------|
| Decision Tree | 88.50 | 0.86 | 0.87 | 0.86 |
| Random Forest | 99.46 | 0.98 | 0.99 | 0.99 |
| Support Vector Machine (SVM) | 92.78 | 0.91 | 0.93 | 0.92 |
| Artificial Neural Network (ANN) | 98.79 | 0.97 | 0.98 | 0.97 |
| Naïve Bayes | 99.46 | 0.99 | 0.99 | 0.99 |
| Long Short-Term Memory (LSTM) | 97.35 | 0.96 | 0.97 | 0.96 |

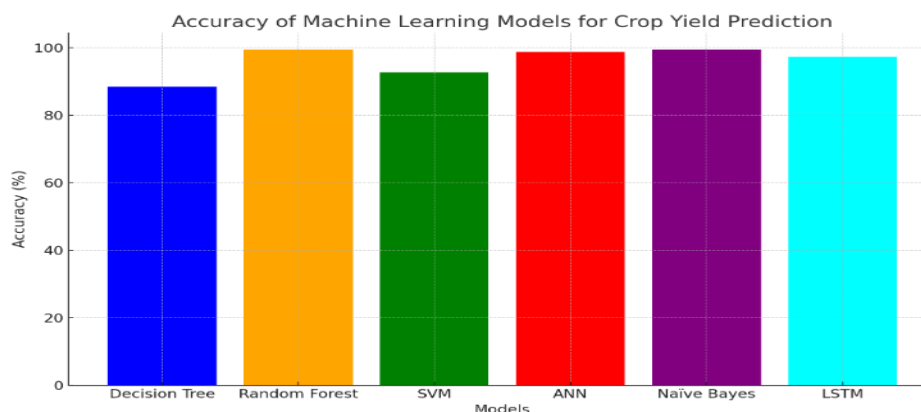


Figure 2: Bar Graph depicting the accuracy of machine learning models for Crop Yield Prediction

5.2 Analysis of Results

Random Forest and Naïve Bayes classifiers attained the best accuracy rate of 99.46% and were, therefore, the best crop-yield-prediction models. ANNs and LSTMs performed well, particularly in dealing with complex temporal agronomic data. Decision Trees and SVMs had low recall and accuracy and may be suffering from potential over fitting or from being vulnerable to fluctuations in a dataset. The findings validate the effectiveness of ensemble methods (Random Forest) and probability classifiers (Naïve Bayes) in dealing with agronomic datasets. Deep learning algorithms (ANN, LSTM) have high flexibility with high computational complexity.

6.0 Real-World Applications

Machine learning has been successfully applied in real farm settings to empower farmers to enhance predictions for harvests, enhance management of resources, and reduce losses in crops. Smart farming systems use machine learning (ML) to make data-driven decisions about crop health, harvesting dates, and irrigation. John Deere integrates ML-driven predictive analysis into its precision agriculture services, boosting productivity and reducing expenditure. IoT and ML integration in crop management transforms monitoring and disease detection, providing early warnings against pests and diseases.

AI-based decision-support systems like Microsoft's AI for Earth and Google's AI-driven agriculture research help farmers plan and forecast yields, particularly in countries affected by climate change. These systems are transforming agriculture and enhancing productivity.

7.0 ML in Crop Yield Prediction: Challenges and Limitations

While ML offers unparalleled advantages in farm forecasting, its use in crop yield estimation has many obstacles and limitations. These obstacles range from limitations in data to computational complexity and real-world adoption challenges.

7.1 Availability and Quality Issues with Data

- Unreliable Collection of Data: Small farmers lack IoT sensors and remote sensing methods and therefore suffer from gaps in data.
- Unstructured and Noisy Data: Datasets in agriculture have inconsistent labels, errors, and missing values, and these can negatively impact model performance.
- Climate Uncertainty: Climate can be unpredictable and change rapidly, and it's hard for ML algorithms to give predictions that can be applied across different locations.

7.2 Computational and Model Limitations

- High Computational Requirement: Deep learning algorithms like CNNs and LSTMs require high computational power, which may be unaffordable to small farms.
- Overfitting Issues: Some ML algorithms perform excellently in training sets but do not work in real-world applications and therefore prove to be unreliable.
- Feature Selection Complexity: Domain knowledge and sophisticated data engineering practices are needed to select high-impact features such as temperature, precipitation, and soil nutrients.

7.3 Real-World Implementation

- Low adoption by farmers: Farmers lack the technical competencies to apply ML-based decision support tools.
- High Implementation Costs: High capital investment in AI-based smart farms with IoT-based monitoring makes it a challenge to implement in developing countries.
- Connectivity problems: Remote farm locations have poor internet connectivity, and it becomes hard to use cloud-based ML.

7.5 Overcoming the Challenge

To enhance the reliability and usability of ML predictions of yield, you can do the following:

- Improving Data Collection Capacity: Greater use of IoT and government-sponsored agriculture databases can bridge gaps in data.
- Enhancing Interpretability: Developing explainable AI (XAI) allows farmers to understand and believe in ML predictions.
- Reduced Computational Costs: Light cloud-based AI algorithms can enable even small farmers to implement ML.

9.0 Conclusion and Future Directions

Machine learning is revolutionizing crop yield forecasting by delivering highly accurate, data-driven predictions that enable farmers to maximize production, allocate resources, and reduce risks. With the application of supervised learning, deep learning, and hybrid models, ML enhances precision agriculture and ensures global food security. Mass adoption, however, is limited by model complexity, data quality,

accessibility, and ethics. In the coming years, a chain of breakthroughs will characterize ML-driven agriculture. Advances in explainable AI (XAI) will improve interpretability in models, and farmers will have trust and comprehension in predictions.

Greater utilization of IoT sensors, block chain, and edge computing will offer real-time, localized decision-making. Lightweight AI models will further make ML-driven farming applications deployable to small-scale farmers, making technological innovation accessible to everyone in agriculture. With ongoing effects from climate change to food production, ML will be critical in reshaping agriculture to respond to uncertain environmental factors.

Future Trends in ML-Based Agriculture

There has been an increase in the use of ML in agriculture, which can be expected to surge with new technologies such as:

- Block chain-based Machine Learning Models for Supply Chain Traceability and Transparency.
- Drone monitoring with AI for mass agriculture.

Quantum machine learning algorithms to improve predictive performance in advanced agriculture systems

Author Contributions

All authors agreed on the content of the study. The author read and approved the final Manuscript.

Funding

External funding was not provided for this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data availability

<https://www.kaggle.com/datasets/mrigaankjaswal/crop-yield-prediction-dataset>

References

- [1] P. Ashok a, B. R. Devi, N. Sharma, M. Behera, A. Gautam, A. Jha, and G. Sinha, "Artificial Intelligence in Water Management for Sustainable Farming: A Review," J. Sci. Res. Rep., vol. 30, no. 6, pp. 511–525, 2024.
- [2] F. Z. Bass in e, T. E. Epule, A. Kechc hour, and A. Cheh bouni, "Recent applications of machine learning, remote sensing, and IoT approaches in yield prediction: a critical review," a r Xiv preprint, arXiv:2306.04566, 2023.

- [3] Y. Chang, J. Latham, M. Licht, and L. Wang, "A data-driven crop model for maize yield prediction," *Commun. Biol.*, vol. 6, no. 1, pp. 1–9, 2023.
- [4] A. Cravero, S. Pardo, P. Galeas, J. LópezFenner, and M. Caniupán, "Data type and data sources for agricultural big data and machine learning," *Sustainability*, vol. 14, no. 23, pp. 1–37, 2022.
- [5] I. Gupta, S. Ayalasomayajula, Y. Shashidhara, A. Kataria, S. Shashidhara, K. Kataria, and A. Undurti, "Innovations in Agricultural Forecasting: A Multivariate Regression Study on Global Crop Yield Prediction," *arXiv preprint, arXiv:2312.02254*, 2023.
- [6] M. JASWAL, "Crop Yield Prediction Dataset," *Kaggle.com*, 2024. [Online]. Available: <https://www.kaggle.com/datasets/mrigaankjaswal/crop-yield-prediction-dataset>. [Accessed: Mar. 4, 2025].
- [7] M. G. J. Kallenberg, B. Maestrini, R. van Bree, P. Ravensbergen, C. Pylianidis, F. van Evert, and I. N. Athanasiadis, "Integrating process-based models and machine learning for crop yield prediction," *arXiv preprint, arXiv:2307.13466*, 2023.
- [8] E. M. B. M. Karunathilake, A. T. Le, S. Heo, Y. S. Chung, and S. Mansoor, "The path to smart farming: Innovations and opportunities in precision agriculture," *Agriculture*, vol. 13, no. 8, pp. 1–26, 2023.
- [9] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, 2021.
- [10] M. Y. Shams, S. A. Gamel, and F. M. Talaat, "Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making," *Neural Comput. Appl.*, vol. 36, no. 11, pp. 5695–5714, 2024.
- [11] J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, S. Sarkar, and A. K. Singh, "Crop yield prediction integrating genotype and weather variables using deep learning," *PLoS One*, vol. 16, no. 6, pp. 1–19, 2021.
- [12] Y. Wang, Q. Zhang, F. Yu, N. Zhang, X. Zhang, Y. Li, M. Wang, and J. Zhang, "Progress in Research on Deep Learning-Based Crop Yield Prediction," *Agronomy*, vol. 14, no. 10, pp. 1–26, 2024.
- [13] J. Zhou, X. Li, B. Liu, B. Wu, and H. Wang, "Progress in research on deep learning-based crop yield prediction," *Agronomy*, vol. 14, no. 10, pp. 1–26, 2022.
- [14] Nazir, T.; Iqbal, M.M.; Jabbar, S.; Hussain, A.; Albathan, M. EfficientPNet—An Optimized and Efficient Deep Learning Approach for Classifying Disease of Potato Plant Leaves. *Agriculture* **2023**, *13*, 841. <https://doi.org/10.3390/agriculture13040841>
- [15] Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Ayari, M.A.; Khan, A.U.; Khan, M.S.; Al-Emadi, N.; Reaz, M.B.I.; Islam, M.T.; Ali, S.H.M. Automatic and Reliable Leaf Disease Detection Using Deep Learning Techniques. *AgriEngineering* **2021**, *3*, 294–312. <https://doi.org/10.3390/agriengineering3020020>
- [16] R. Kumar, A. Chug, A. P. Singh, and D. Singh, "A Systematic Analysis of Machine Learning and Deep Learning Based Approaches for Plant Leaf Disease Classification: A Review," *Journal of Sensors*, vol. 2022, pp. 1–13, Jul. 2022, doi: 10.1155/2022/3287561.
- [17] Shafik, W., Tufail, A., De Silva Liyanage, C. et al. Using transfer learning-based plant disease classification and detection for sustainable agriculture. *BMC Plant Biol* **24**, 136 (2024). <https://doi.org/10.1186/s12870-024-04825-y>
- [18] Srivastava, Animesh, et al. "Potato Leaf Disease Detection Method is Based on the Support Vector Machines." 2024 Second International Conference on Advanced Computing & Communication Technologies (ICACCTech). IEEE, 2024.
- [19] Kayal, Mrinmoy, MohinikantaSahoo, and JayadeepPati. "Review on Mental Healthcare System Using Data Analytics and IoT." International Conference on Global Mental Health and Public Health Challenges and Innovation. Cham: Springer Nature Switzerland, 2022.

- [20] Kayal, Mrinmoy, et al. "Quantum-Inspired Aspect-Based Sentiment Analysis Using Natural Language Processing." *Advances in Quantum Inspired Artificial Intelligence*. Springer, Cham, 2025. 151-169.
- [21] Chaudhari , Chetan, Sapana Fegade, SasankoSekharGantayat, KumariJugnu, and VikashSawan. "Influenza Diagnosis Deep Learning: Machine Learning Approach for Pharyngeal Image Infection." (2024)