

**INDIVIDUAL ASSIGNMENT**

**TECHNOLOGY PARK MALAYSIA**

**CT098-3-2-RMCT**

**RESEARCH METHODS FOR COMPUTING AND TECHNOLOGY**

**APU2F2411CS(AI)**

**HANDOUT DATE:**

**HAND IN DATE:**

**LECTURER NAME:**

**STUDENT NAME:**

**TP NUMBER:**

**FRIDAY, 20TH JUNE 2025**

**FRIDAY, 1ST AUGUST 2025**

**DR. MURUGANANTHAN VELAYUTHAM**

**MANREEN KAUR A/P JAGJIT SINGH**

**071290**

Table of Contents

[**1.** **Introduction** 3](#_Toc204959390)

[**2.** **Literature Review** 4](#_Toc204959391)

[**3.** **Methodology** 6](#_Toc204959392)

[**4.** **Conclusion** 7](#_Toc204959393)

[**5.** **References** 8](#_Toc204959394)

**Development of an Agri Crop Yield Prediction Model**

Manreen Kaur A/P Jagjit Singh

TP071290@mail.apu.edu.my

* + - 1. **Introduction**

Building on the ideas introduced in Part 1, this section delves into the use of machine learning (ML) to improve the accuracy of agricultural yield predictions, a key aspect of advancing modern farming. Previously, the discussion centered around the lack of sufficient agricultural data, a major barrier in creating reliable prediction models based on soil nutrients and environmental variables. With growing concerns over food security and the increasing demand for sustainable agriculture, ML offers practical tools to boost crop yields while supporting eco-friendly farming methods.

In this stage of the research, various Machine Learning algorithms are systematically evaluated to determine their performance, practicality, and adaptability in agricultural yield prediction. The algorithms considered span both classical and contemporary techniques, including Decision Trees (DT), K-Nearest Neighbors (KNN), Random Forests (RF), Support Vector Machines (SVM), Extreme Gradient Boosting (XGB), and Long Short-Term Memory (LSTM) networks. Their effectiveness is assessed using core agricultural indicators such as soil nutrients—Nitrogen (N), Phosphorus (P), and Potassium (K), as well as meteorological variables including rainfall, humidity, and temperature.

Recent studies highlight a growing trend toward the use of ensemble methods and deep learning in agricultural applications, largely due to their robustness and predictive accuracy in handling complex and non-linear datasets. Algorithms like XGBoost and Random Forests consistently perform well, particularly in high-dimensional environments, while maintaining resistance to overfitting. Meanwhile, Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, are especially effective in modeling time-series data, allowing them to track long-term trends such as crop performance across multiple growing seasons. This progression reflects how rapidly the field is evolving with the help of advanced machine learning tools.

While these models offer promising results, they’re not without flaws. Issues like imbalanced datasets, unpredictable regional differences, and a lack of transparency in how the models make decisions. To address these challenges, many suggest integrating real-time data from IoT sensors or combining multiple algorithms to create more reliable and adaptable systems. Since agricultural needs vary widely across regions, blending technologies such as pairing lightweight edge devices like Raspberry Pi with cloud services has become a practical approach for building smarter, and more cost-effective solutions in the field.

In the implementation part, this research uses Python as the development environment to build, train, and test the models. The methodology employs a systematic pipeline that begins with data preprocessing, which includes normalization and standardization, to prepare the dataset for optimal machine learning performance. Key agricultural features are retrieved and adjusted to enhance model learning. The chosen algorithms are then configured and tested against performance criteria like accuracy, F1-score, and mean squared error (MSE).

The comparative examination of DT, RF, KNN, XGB, and SVM seeks to establish which classical and ensemble models provide the best balance of accuracy, interpretability, and scalability. These machine learning models are compared to standard yield estimation approaches to demonstrate gains in prediction accuracy and adaptability. In contrast, LSTM is evaluated independently because of its distinct structure as a deep learning model and its ability to capture time-dependent patterns, which are especially important for long-term agricultural forecasting.

By comparing various algorithms, this work sheds light on how different machine learning methodologies can be used to accurately estimate crop yields under varying agricultural settings. The study not only shows each model's predictive strengths, but also provides suggestions on how to apply them in practice, particularly in contexts with limited technological resources.

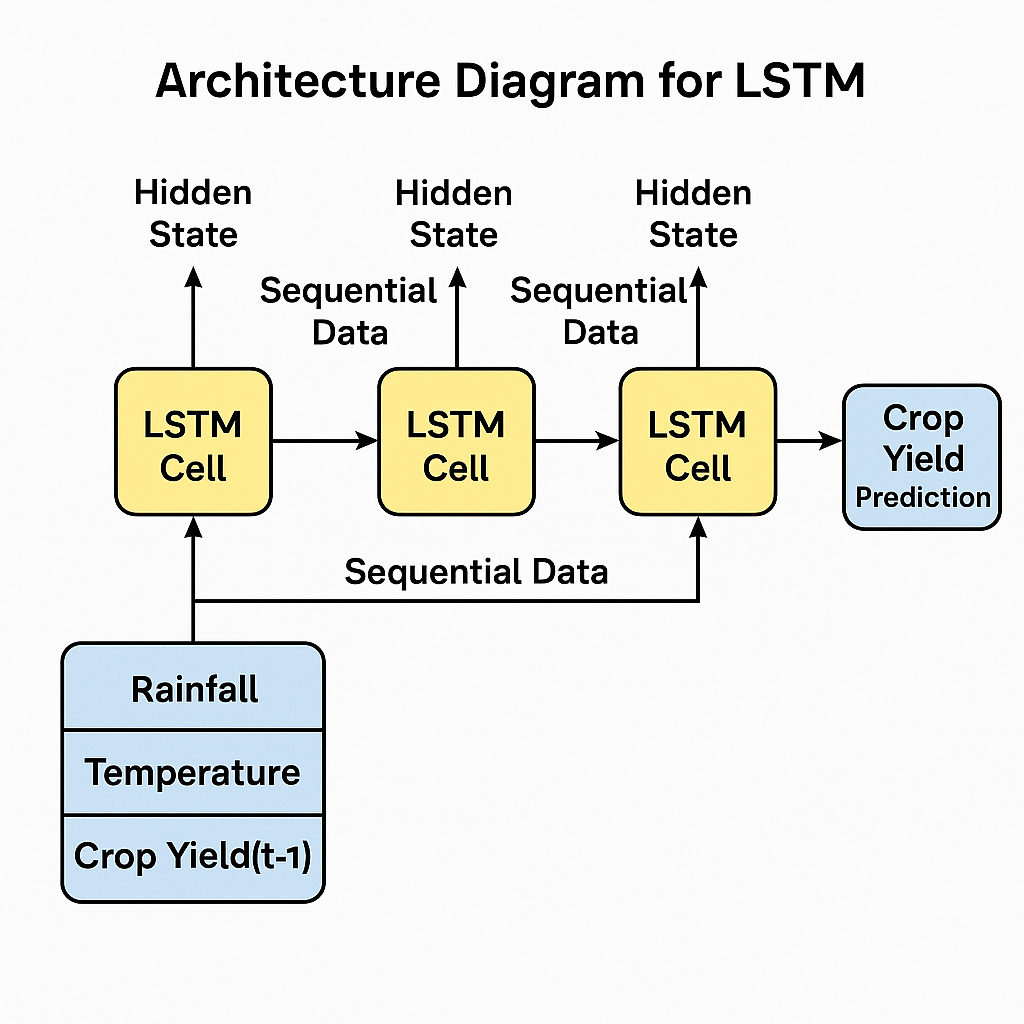
* + - 1. **Literature Review**

2.1 Domain 1

Deep learning models are prominent in agricultural production prediction because of their capacity to model nonlinear relationships and extract complicated patterns from input data. Long Short-Term Memory (LSTM) networks, a form of Recurrent Neural Network (RNN), are well-known for their ability to handle sequential inputs and temporal relationships, which is useful when assessing seasonal or time-series agricultural data such as rainfall and temperature trends.

The article "Automated Yield Prediction Using LSTMs and Remote Sensing Data" (2025) shows how deep learning can outperform classical models in terms of forecast accuracy, especially when combining previous vegetation indices and meteorological data. While this study predominantly uses remote sensing inputs, the LSTM technique can also be used to structured environmental datasets such as those employed in this study like soil nutrient levels and weather patterns.

Although the current study focuses on classic machine learning methods (DT, RF, KNN, SVM, and XGB), Part 2 introduces LSTM as an advanced comparison approach. Its inclusion is justified by its ability to represent temporal variability in environmental parameters, which provides insight into long-term crop response under changing conditions. However, deep learning models have disadvantages, such as high data requirements and interpretability concerns, which are addressed in this study.



**2.2 Domain 2**

The use of cloud-based IoT technologies in agriculture has greatly improved data collection and decision-making processes. The review article "Smart IoT Cloud-based Intelligent System for Crop Yield Prediction" (2025) looks at existing techniques that use sensor networks, weather stations, and cloud infrastructure to monitor critical metrics like soil moisture, temperature, and crop health. It emphasizes the advantages of cloud platforms for scalability, real-time analytics, and intelligent automation. Fog computing is cited in the literature as a possible intermediary layer for reducing latency between edge devices and cloud services. Furthermore, the study demonstrates that small-scale farmers can profit from edge-cloud hybrid systems, despite challenges such as cost and poor internet connectivity.

While fog computing is mentioned in the studied literature, it is not part of the suggested implementation for this study. Instead, the system takes a direct edge-to-cloud integration strategy, connecting field-deployed sensors and microcontrollers to cloud platforms to enable real-time data transmission and processing. This design decision provides simplicity, cost-effectiveness, and compatibility with the technological constraints common in small-scale farming situations. However, future developments may include fog computing to boost processing efficiency and reduce network dependency in latency-sensitive circumstances.

A diagram of a cloud platform

AI-generated content may be incorrect.

**2.3 Domain 3**

Machine learning and deep learning algorithms have undergone significant testing for yield prediction. In "Crop Yield Prediction in Agriculture: A Comprehensive Review of ML and DL Approaches with Insights for Future Research and Sustainability" (2024), the authors evaluate models such as Random Forests, XGBoost, CNNs, and DNNs using datasets from India, Brazil, and Europe. When big labeled datasets are available, deep learning models consistently outperform classical machine learning algorithms in terms of accuracy and scalability. However, the report highlights overfitting, interpretability, and a lack of data availability as recurring limitations in deep learning applications. It proposes ensemble methods and transfer learning as future approaches to bridging performance and generalization.

**2.4 Similar system**

Table 1 summarizes the comparisons between similar systems. Before presenting the table, a brief summary of the three chosen systems is provided below. Several recent systems show advances in multimodal data fusion and hybrid learning for crop forecasting. Wang et al. (2024) propose a taxonomy of deep learning architectures, including CNNs, DNNs, and attention-based models designed for agricultural yield prediction. They present a performance comparison and dataset analysis, emphasizing the need of cross-domain adaptability.   
  
Suresh et al. (2023) undertake a critical review of systems that combine remote sensing, IoT, and machine learning. Their findings indicate that combining heterogeneous data (for example, SAR imaging and real-time sensors) with ensemble models considerably increases yield prediction accuracy and resource management.

Similarly, Dagadu Yewle et al. (2025) propose RicEns-Net, a deep ensemble architecture for combining SAR, multispectral, and meteorological data. The system delivers cutting-edge yield prediction performance on benchmark datasets, emphasizing the significance of feature fusion and cross-modal learning.   
  
These systems contribute to a developing trend of strong, AI-powered yield prediction platforms, but there are still common gaps in data availability, cost-efficiency for small-scale farms, and real-world deployment.   
  
The comparison of similar systems is shown in Table 1. Three alternative systems were examined, and each system was reported in three separate articles, as indicated in Table 1.

Table 1 Similar system comparisons with research gaps identified

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Journal (Year)** | **Authors** | **Title** | **Key Findings** | **Research Gaps** |
| Agronomy (Wang et al., 2024) | Wang et al. | |  | | --- | |  |  |  | | --- | | Progress in Research on Deep Learning‑Based Crop Yield Prediction | | Comprehensive taxonomy and analysis of Deep Learning models (CNN, LSTM, attention) for crop yield | Limited exploration of generalizability across regions and crops |
| arXiv (Suresh et al., 2023) | Suresh, Reddy, & Kumar | Recent Applications of Machine Learning, Remote Sensing, and IoT in Yield Prediction | Remote sensing, Machine Learning, IoT integration shows high accuracy in smart farming | Cost, scalability, and deployment for small farms are not addressed |
| arXiv (Dagadu Yewle et al., 2025) | Dagadu Yewle et al. | Multi-modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction | Proposed RicEns-Net outperforms existing models through feature fusion | Limited real-time implementation and system adaptability |

* + - 1. **Methodology**

A diagram of a model training

AI-generated content may be incorrect.

**3.1 Target user**

The main users of this research are farmers, agricultural planners, agronomists, and government organizations in charge of food security and agricultural policymaking. These stakeholders rely on accurate agricultural production projections to help them allocate resources, plan planting tactics, and mitigate risks. The suggested approach also helps small-scale farmers who do not have access to complex forecasting tools by providing a data-driven decision-support model.

**3.2 Sampling method**

A purposive sample strategy is used to choose agricultural datasets containing crop yield data, soil nutrients (Nitrogen, Phosphorus, Potassium), and environmental factors (temperature, humidity, rainfall). Datasets are chosen based on their completeness, quality, and suitability for supervised machine learning models. Benchmark datasets from open agricultural sources and research institutes are utilized to assure uniformity and reproducibility.   
  
Furthermore, models such as Decision Trees (DT), Random Forests (RF), K-Nearest Neighbors (KNN), XGBoost (XGB), Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) are sampled and chosen to examine and compare performance across several machine learning methodologies.

**3.3 Data collection method**

This study uses a quantitative approach, relying on publicly available secondary datasets to assist statistical analysis and predictive modeling. The key data sources include government-released crop yield records from India, agricultural repositories from the Food and Agriculture Organization (FAO), and curated datasets from platforms like Kaggle. Furthermore, open-source IoT sensor data on climate conditions and soil properties is used to capture environmental variability that affects crop production results.

Following data capture, the gathering process involves several crucial steps. Data preprocessing begins with the treatment of missing values, followed by normalization and standardization to maintain consistency across different data scales. The major variables that determine crop output are then identified via feature extraction, including soil nutrient levels (Nitrogen, Phosphorus, Potassium) and weather-related parameters such as rainfall, humidity, and temperature. The revised dataset is then used to train and test several machine learning and deep learning models constructed in Python with libraries like Scikit-learn, TensorFlow, and Keras.

Models are evaluated using performance metrics such as accuracy, F1-score, and Mean Squared Error (MSE), which provide detailed information about each algorithm's predicting capabilities. This methodological approach is well-suited to the study's goal of yield forecasting since it allows for the use of big numerical datasets and stresses model performance using quantitative metrics.

The use of machine learning approaches is supported by their potential to deliver scalable, flexible, and data-driven solutions to agricultural forecasting problems. Unlike traditional estimating techniques, which frequently rely on static rules and expert opinion, machine learning algorithms may learn from past experiences and generalize across different situations. The integration of Long Short-Term Memory (LSTM) networks contributes to this goal by allowing for the analysis of time-dependent trends, which provides deeper insights into crop performance throughout many seasons.

Because the research is entirely based on open-source, secondary datasets, there is no involvement of human participants, and the accompanying ethical implications are minor. Nonetheless, ethical issues are crucial. These include properly citing data sources, protecting the privacy of any potentially sensitive information, and pushing for equitable access to developed prediction models, particularly for smallholder and marginalized farming communities. This study does not have any safety issues because there is no physical fieldwork.

Certain shortcomings are mentioned in this study. These include regional data imbalance, inadequate records that may influence model accuracy, and the possibility of overfitting, especially with more complex models like LSTM. To address these problems, methods such as stratified sampling and k-fold cross-validation are used to improve model reliability. Furthermore, ensemble and hybrid learning techniques are believed to improve generalizability across various agricultural situations. Data augmentation and targeted feature selection help to improve the model's predictive capability.

* + - 1. **Conclusion**

This study contributes to the continuing endeavor to incorporate machine learning into agriculture by providing a comparative analysis of multiple ML and DL models for agricultural production prediction. Building on the basic goals stated in Part 1, the study expands the scope by using deep learning models such as LSTM and evaluating their applicability in time-series agricultural data. The research intends to discover effective algorithms capable of forecasting yields under a variety of environmental and soil conditions using a structured methodology that includes public datasets, IoT sensor inputs, and performance-driven model evaluation.

The literature study reinforces the increased reliance on ensemble and hybrid learning models, while also highlighting practical issues such as data availability, scalability, and real-world deployment limits. The technique tackles these concerns by carefully selecting datasets, benchmarking performance, and implementing ethical precautions.  
  
Finally, our study provides a solid foundation for constructing an intelligent, scalable agricultural production forecast system that can serve both large-scale planners and resource-constrained smallholder farmers. It adds to the corpus of knowledge by providing not only comparative insights into algorithm performance, but also practical advice for implementation in a variety of agricultural settings.

* + - 1. **References**

Abu, J. M., & Azmi. (2024). Crop yield prediction in agriculture: a comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability - Universiti Putra Malaysia Institutional Repository. Upm.edu.my.

<http://psasir.upm.edu.my/id/eprint/116081/1/116081.pdf>

B. M. Beena, Dr., Manju Khanna, Dr., & Shalini G. (2025). IoT‑Cloud based intelligent system for crop yield prediction: A review. ResearchGate.

<https://www.researchgate.net/publication/390427177_IoT-Cloud_based_Intelligent_System_for_Crop_Yield_Prediction_A_Review>

Brahmeshwara Sr., K. B. (2025). Automated yield prediction using LSTMs and remote sensing data. ResearchGate.

<https://www.researchgate.net/publication/390164176_AUTOMATED_YIELD_PREDICTION_USING_LSTMS_AND_REMOTE_SENSING_DATA‌>

Kalyani, Y., & Collier, R. (2021). A Systematic Survey on the Role of Cloud, Fog, and Edge Computing Combination in Smart Agriculture. *Sensors*, *21*(17), 5922. <https://doi.org/10.3390/s21175922>

Marković, D., Stamenković, Z., Đorđević, B., & Ranđić, S. (2024). Image Processing for Smart Agriculture Applications Using Cloud-Fog Computing. Sensors, 24(18), 5965. <https://doi.org/10.3390/s24185965>

Suresh, P. R., Reddy, K. S., & Kumar, P. R. (2023). Recent applications of machine learning, remote sensing, and IoT approaches in yield prediction: A critical review. arXiv. <https://arxiv.org/abs/2306.04566>

Wang, Y., Zhang, Q., Yu, F., Zhang, N., Zhang, X., Li, Y., Wang, M., & Zhang, J. (2024). Progress in Research on Deep Learning-Based Crop Yield Prediction. (Agronomy)

<https://www.researchgate.net/publication/384546942_Progress_in_Research_on_Deep_Learning-Based_Crop_Yield_Prediction>

Yewle (2025). Multi-modal data fusion and deep ensemble learning for accurate crop yield prediction. arXiv.

<https://arxiv.org/pdf/2502.06062>