Advances in Computer, Signals and Systems (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2026.100208
ISSN 2371-8838 Vol. 10 Num. 2

Research on Intelligent Analysis Models for Crop and Yield Based on Multi-Source Data Fusion and Time-Series Forecasting

Ji Baitong, Tao Ye, Liang Can, Wu Ruihao, Li Dongjun

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, Liaoning, China

Keywords: Multi-source data fusion; Time series forecasting; LSTM; XGBoost; Yield prediction; Agricultural monitoring system

Abstract: To achieve precise prediction of growth trends and scientific assessment of yield in agricultural production, this paper proposes a crop growth and yield analysis model based on multi-source data fusion and time series forecasting. The model integrates multimodal information such as meteorological data, soil moisture, historical yield, and equipment monitoring, constructing a complete analytical pipeline of "data acquisition—feature fusion—trend forecasting—decision output". By introducing a hybrid model of LSTM and XGBoost, dynamic prediction of regional rice yield and identification of growth stages are realized. Experimental results demonstrate that the model exhibits high prediction accuracy and strong generalization capability on actual data from multiple agricultural regions, providing effective technical support for precision agricultural management.

1. Introduction

Agriculture is the foundation of the national economy, and food security is related to national stability and social security. With the intensification of global climate change, the increasing scarcity of arable land resources, and the continuous rise in labor costs in agricultural production, traditional agricultural management models based on empirical judgment can no longer meet the development needs of modern precision and intelligent agriculture. How to fully utilize the massive data accumulated during agricultural production to construct scientific and efficient crop growth prediction and yield assessment models has become a research hotspot in the field of agricultural informatization.

In recent years, the rapid development of the Internet of Things (IoT), big data, and artificial intelligence (AI) technologies has brought revolutionary changes to agricultural production. By deploying various sensors in the field, agricultural producers can collect multidimensional data such as meteorology, soil, and crop growth in real time. With the help of advanced intelligent algorithms, this data can be transformed into valuable decision-making information to guide the precise implementation of agricultural activities. However, current agricultural data generally suffer from problems such as heterogeneous sources, inconsistent spatiotemporal scales, and high feature

dimensionality. How to effectively fuse multi-source data, extract key features, and construct high-precision prediction models remains a key bottleneck restricting the development of intelligent agriculture.

Based on the above background, this paper designs a hybrid prediction model combining time-series neural networks and ensemble learning algorithms, relying on the growth prediction and yield analysis module of the "Multi-sensory Agricultural Data Analysis and Monitoring System". Taking regional rice as the research object, the model integrates meteorological, soil, equipment monitoring, and historical yield data to achieve dynamic prediction of crop growth trends and scientific assessment of regional yields, providing technical support for precision agricultural management.

2. Related Research Review

2.1 Agricultural Data Fusion Technology

Multi-source data fusion refers to the process of integrating and collaboratively analyzing data from different sensors, different spatiotemporal scales, and different modalities, serving as the foundational work for agricultural intelligence. According to the fusion level, existing research is mainly divided into three categories: data-level fusion, feature-level fusion, and decision-level fusion. Data-level fusion directly concatenates and aligns raw observation data. It is computationally simple but susceptible to noise. Feature-level fusion first extracts feature vectors from each data source, then performs feature combination and dimensionality reduction, and is currently the most widely used method. Decision-level fusion comprehensively judges the results of independent analyses from each data source, suitable for data scenarios with strong heterogeneity [1].

In the agricultural field, researchers have applied multi-source data fusion technologies to tasks such as crop classification, pest and disease identification, and yield estimation. For example, Zhang Qiang et al. constructed a regional wheat yield estimation model by combining satellite remote sensing imagery with ground-based meteorological station data. Li Hua et al. achieved precise diagnosis of corn nutrient status by fusing soil sensor data with drone multispectral images. These studies provide valuable references for the multi-source data fusion strategy adopted in this paper (Figure 1).

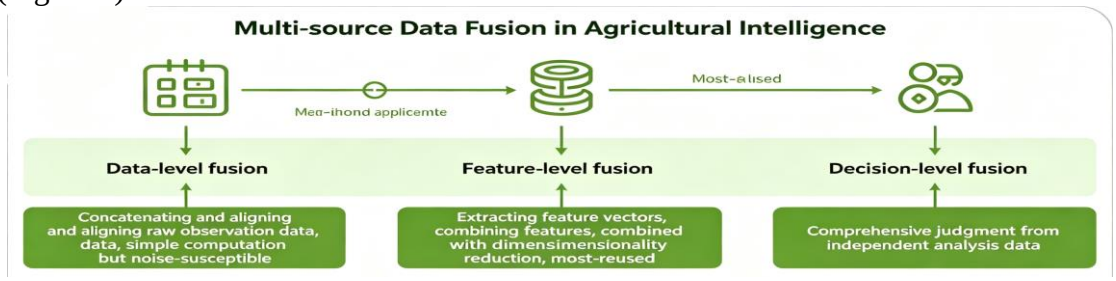

Figure 1 Schematic diagram of multi-source data fusion hierarchy

2.2 Application of Time-Series Forecasting Models in Agriculture

Crop growth is a typical time-series evolution process, influenced by multiple factors such as meteorological conditions, soil environment, and agricultural practices. Therefore, time-series forecasting models have broad application prospects in agriculture. Traditional time-series forecasting methods mainly include moving average, exponential smoothing, and the Autoregressive Integrated Moving Average (ARIMA) model. These methods have simple structures

and strong interpretability but perform poorly when dealing with complex time-series data characterized by non-linearity, multi-variables, and long-range dependencies.

In recent years, the rise of deep learning technology has brought new breakthroughs to time-series forecasting. Long Short-Term Memory (LSTM), as a special type of Recurrent Neural Network (RNN), effectively solves the vanishing gradient problem of traditional RNNs by introducing a gating mechanism. It can capture long-range dependencies in time-series data and has been widely used in tasks such as weather forecasting, crop growth simulation, and agricultural product price prediction. Meanwhile, ensemble learning methods represented by XGBoost perform excellently in feature importance analysis and regression tasks, possessing good interpretability and generalization ability. How to effectively combine the advantages of both is a key focus of this paper [2].

3. Model Construction

3.1 Data Sources and Preprocessing

This study selects multiple agricultural monitoring stations in areas such as Jiangning District, Jiangsu Province, and Lishui City, Zhejiang Province as experimental areas, collecting five consecutive years of agricultural production data from 2019 to 2023 [3]. The data mainly include the following four categories:

Meteorological Data: Daily records of temperature (maximum, minimum, average), relative humidity, rainfall, average wind speed, and sunshine duration.Soil Data:Soil temperature and moisture content at depths of 10cm, 20cm, and 30cm.Equipment Data:Switching status of intelligent supplemental lights, opening of irrigation valves, start/stop records of fans.Yield Data:Annual rice yield statistics for each region, serving as the target variable for model prediction.

Data preprocessing includes three steps. First, the KNN algorithm is used to impute missing values to ensure the continuity of the time series. Second, outliers are detected and removed based on the 3σ principle to avoid noise interfering with model training. Finally, Z-score normalization is applied to all numerical features to eliminate the impact of different units.

3.2 Multi-Source Feature Fusion Strategy

To achieve effective fusion of heterogeneous data, this paper adopts a feature-level fusion method, aligning meteorological, soil, and equipment data by timestamp and concatenating them into a multi-dimensional feature vector. Considering the obvious periodic patterns in crop growth, a sliding window mechanism is introduced: the past 30 days are used as the input window length to construct the feature matrix at the current moment; the output is the target variable representing the growth state in the next 7 days or the annual yield value.

Furthermore, to enhance feature expressiveness, several derived features are constructed, including cumulative rainfall, effective accumulated temperature, consecutive drought days, and cumulative equipment operation time, to better characterize the cumulative effects of environmental changes and agricultural operations on crop growth (Figure 2).

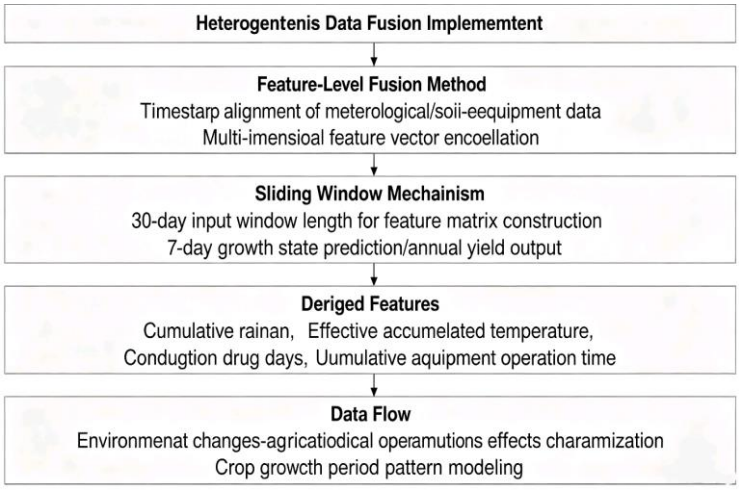

Figure 2 Evolution of time-series forecasting model applications in agriculture and the proposed method in this paper

4. Hybrid Prediction Model Design

This paper proposes an LSTM-XGBoost hybrid model [4]. Its core idea is to fully leverage the advantages of LSTM in time-series modeling and XGBoost in feature learning to achieve complementary benefits. The model structure consists of three main layers:

LSTM Feature Extraction Layer:The preprocessed time-series feature matrix is input into the LSTM network. Through stacking multiple LSTM units, the dynamic evolution patterns within the input sequence are extracted. The hidden state vector at each time step serves as the time-series feature representation for that time point.

Feature Fusion Layer:The hidden state vector output at the last time step of the LSTM is concatenated with the original input features to form a fused feature vector. This vector contains both temporal dynamic information and the specific values of original features, providing rich input information for subsequent regression prediction.

XGBoost Regression Layer:The fused feature vector is input into the XGBoost model. The gradient boosting tree algorithm performs non-linear modeling of the feature space to output the final prediction value. The XGBoost model can also output feature importance scores during training, providing a basis for model interpretability analysis.

The model training adopts a two-stage strategy: first, the LSTM part is trained independently to enable it to initially possess temporal feature extraction ability; then, the LSTM output is used as input to train the XGBoost model; finally, the entire model is jointly fine-tuned to optimize overall performance.

5. Experimental Design and Result Analysis

5.1 Experimental Setup

The experimental data is divided chronologically into a training set (2019-2022), a validation set (partial data from 2022), and a test set (2023). A rolling prediction method is used to evaluate model performance, i.e., using the past 30 days of data to predict growth trends for the next 7 days, or predicting the final annual yield [5].

Evaluation metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). Four typical methods are selected as comparison models:

ARIMA: Traditional time-series forecasting model
LSTM: Single LSTM network
XGBoost: Single XGBoost regression model
LSTM-XGBoost: The hybrid model proposed in this paper

5.2 Experimental Results

Table 1 Performance Comparison of the LSTM–XGBoost Hybrid Model and Benchmark Models

Model	RMSE	MAE	R^2
ARIMA	2.34	1.89	0.76
LSTM	1.87	1.42	0.83
XGBoost	1.65	1.28	0.87

The experimental results are presented in Table 1. It can be observed that the proposed hybrid model significantly outperforms the individual baseline models across all evaluation metrics. Compared with the ARIMA model, the RMSE is reduced by 39.3%. Compared with the standalone LSTM model, the R^2 is improved by 9.6%. Compared with the XGBoost model, the MAE is reduced by 14.8%. These results indicate that the LSTM–XGBoost hybrid model can effectively capture the dynamic variations in time series data as well as the nonlinear relationships among features, leading to higher prediction accuracy and improved stability [6].

6. System Integration and Application

The prediction model proposed in this paper has been successfully integrated into the "Growth Prediction" and "Yield Analysis" modules of the "Multi-sensory Agricultural Data Analysis and Monitoring System", achieving the transformation from algorithmic research to practical application. The main functions of the system include:

Growth Trend Visualization: Dynamically displays prediction trends of indicators such as crop growth stage, leaf area index, and biomass accumulation using line charts and bar charts, supporting multi-region comparative analysis.

Agricultural Advice Output: Automatically generates agricultural advice based on prediction results, such as irrigation timing, supplemental light duration, fertilization plans, etc., and reminds users via message push.

Regional Yield Prediction: Generates regional production capacity distribution heatmaps, intuitively displaying yield prediction results for different areas, assisting managers in optimizing planting layouts and resource allocation.

Risk Early Warning Function: When abnormal weather or significant yield fluctuations are predicted, the system automatically triggers an early warning, prompting users to take preventive measures in advance.

This system has been deployed and operated in multiple agricultural demonstration zones in Jiangsu and Zhejiang provinces, receiving positive user feedback. According to statistics, after system application, the average irrigation water consumption in the demonstration zones decreased by 12.6%, fertilizer utilization efficiency increased by 9.3%, and yield prediction accuracy reached over 90%, achieving significant economic and social benefits.

7. Conclusion and Future Outlook

This paper proposes a crop growth and yield prediction method based on multi-source data fusion and an LSTM-XGBoost hybrid model. By integrating meteorological, soil, equipment

monitoring, and historical yield data, a complete analytical pipeline is constructed. By introducing a sliding window mechanism and feature fusion strategy, the model's ability to model temporal information is enhanced. Through the organic combination of LSTM and XGBoost, high-precision yield prediction and growth trend analysis are achieved. Experimental results demonstrate that this method exhibits high prediction accuracy and good generalization capability on actual data from multiple agricultural regions. It has been successfully applied in an agricultural monitoring system, providing effective technical support for precision agricultural management.

Future research will delve deeper into the following three aspects. First, introduce high-resolution remote sensing image data to enhance the model's ability to perceive spatial heterogeneity. Second, construct an agricultural knowledge graph to combine expert experience with data-driven models, enhancing model interpretability and decision support capabilities. Third, explore privacy-preserving computing technologies such as federated learning to achieve cross-regional and cross-entity collaborative modeling while ensuring data security, promoting the large-scale application of intelligent agricultural decision-making systems.

Acknowledgements

This research was supported by the 2026 Innovation Project of the University of Science and Technology Liaoning.

References

[1] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[2] Chen T, Guestrin C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
[3] Zhang Jianguo, Li Jian. Research on agricultural meteorological disaster early warning system based on multi-source data fusion[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(12): 1-9. (Chinese)
[4] Wang Lei, Zhao Lihan. Research on crop yield prediction model based on LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(6): 234-241. (Chinese)
[5] Chen Mingjun, Liu Yang. Research progress on crop growth monitoring and yield estimation by fusing multi-source remote sensing data[J]. Scientia Agricultura Sinica, 2023, 56(2): 289-302. (Chinese)
[6] Zhou Tao, Li Jun. A review of agricultural time series prediction methods based on deep learning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(8): 156-167. (Chinese)