

## Article

# Pest and Disease Prediction and Management for Sugarcane Using a Hybrid Autoregressive Integrated Moving Average—A Long Short-Term Memory Model

Minghui Wang <sup>1,2</sup>  and Tong Li <sup>2,3,\*</sup><sup>1</sup> College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming 650201, China; p2022034@ynau.edu.cn<sup>2</sup> The Key Laboratory for Crop Production and Smart Agriculture of Yunnan Province, Kunming 650201, China<sup>3</sup> Big Data College, Yunnan Agricultural University, Kunming 650201, China

\* Correspondence: tli@ynu.edu.cn

**Abstract:** This study introduces a hybrid AutoRegressive Integrated Moving Average (ARIMA)—Long Short-Term Memory (LSTM) model for predicting and managing sugarcane pests and diseases, leveraging big data for enhanced accuracy. The ARIMA component efficiently captures linear patterns in time-series data, while the LSTM model identifies complex nonlinear dependencies. By integrating these two approaches, the hybrid model effectively handles both linear trends and nonlinear fluctuations, improving predictive performance over conventional models. The model was trained on 33 years of meteorological and pest occurrence data, and its effectiveness was evaluated using mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE). The results show that the ARIMA-LSTM model achieves an MSE of 2.66, RMSE of 1.63, and MAE of 1.34, outperforming both the standalone ARIMA model (MSE = 4.97, RMSE = 2.29, MAE = 1.79) and LSTM model (MSE = 3.77, RMSE = 1.86, MAE = 1.45). This superior performance highlights its ability to effectively capture seasonal variations and complex nonlinear patterns in pest outbreaks. Beyond accurate forecasting, this model provides valuable decision-making support for agricultural management, aiding in early intervention strategies. Future enhancements, including the integration of additional variables and climate change factors, could further expand its applicability across diverse agricultural sectors, improving crop yield stability and pest control strategies in an increasingly unpredictable climate.

**Keywords:** pest and disease; prediction and management; model optimization; ARIMA-LSTM



Academic Editor: Thomas Thomidis

Received: 8 January 2025

Revised: 13 February 2025

Accepted: 19 February 2025

Published: 26 February 2025

Corrected: 3 April 2025

**Citation:** Wang, M.; Li, T. Pest and Disease Prediction and Management for Sugarcane Using a Hybrid Autoregressive Integrated Moving Average—A Long Short-Term Memory Model. *Agriculture* **2025**, *15*, 500. <https://doi.org/10.3390/agriculture15050500>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The challenges facing agricultural production are increasing with global climate change, the expansion of agricultural production and the increased complexity of pests and diseases. Accurate prediction and management of pests and diseases has become a key issue in ensuring food security and sustainable agricultural development. Sugarcane is one of the most important commercial crops globally, contributing significantly to the economies of major producing countries such as Brazil, India, China, and Thailand. According to the FAO, global sugarcane production exceeds 1.9 billion tons annually, and the industry supports millions of livelihoods. However, pests and diseases such as sugarcane borers, smut, and rust have led to substantial yield losses, often exceeding 20% in severely affected regions, translating into billions of dollars in economic losses. Thus, accurate

prediction and management of sugarcane pests and diseases are critical for ensuring global food security, sustaining agricultural productivity, and protecting economic interests [1,2]. Traditional statistical models such as ARIMA are well-established for time series forecasting due to their ability to capture linear trends and seasonal variations. However, agricultural pest and disease data exhibit highly nonlinear and dynamic characteristics influenced by multiple exogenous variables such as temperature, humidity, and rainfall. These factors introduce complex dependencies that linear models alone fail to capture, leading to suboptimal predictions. Machine learning models, particularly deep learning-based approaches like Long Short-Term Memory (LSTM) networks, have demonstrated superior performance in handling long-term dependencies and nonlinear data structures. By integrating ARIMA and LSTM into a hybrid framework, we can leverage ARIMA's strength in linear trend detection and LSTM's capability in nonlinear feature extraction, significantly enhancing predictive accuracy and robustness [3]. Consequently, there is a pressing requirement for more precise models that offer enhanced predictive capabilities to assist farmers and agricultural managers in addressing these issues [4].

In recent times, the advancement of big data and artificial intelligence (AI) technologies has led to notable advancements in the prediction of agricultural pests and diseases. Conventional time series forecasting techniques, including the autoregressive integrated moving average (ARIMA) model, have been extensively applied for forecasting agricultural yields as well as monitoring pests and diseases [3]. The ARIMA model is particularly effective in managing linear datasets and is well-suited for identifying long-term trends and seasonal variations in time series data [5]. However, due to the complexity and dynamics of agricultural production data, it is difficult for a single linear model to cope with the challenges of nonlinearity and long- and short-term dependencies. Therefore, scholars have begun to combine traditional statistical models with deep learning models to form hybrid models to improve prediction performance. Thus, hybrid models of big data and machine learning are gradually showing [6–8] strong advantages in agricultural pest and disease prediction. In particular, hybrid models combining ARIMA with Long Short-Term Memory (LSTM) neural networks excel in handling both linear and nonlinear data [9]. For example, ARIMA models are able to capture linear trends in time series, while LSTMs are able to effectively deal with nonlinear features and long- and short-term dependencies in serial data, thereby improving the accuracy of forecasts [1,10–12]. In research on agricultural pests and diseases, a variety of hybrid models have been applied to different crop and pest prediction scenarios. For example, Guo et al. [13]. In their prediction of the incidence of hepatitis B, verified that a model combining ARIMA and LSTM can effectively cope with seasonal fluctuations and substantially improve prediction accuracy. Similarly, Dhawan demonstrated the significant advantage of hybrid models in nonlinear data processing in a prediction study of sugarcane disease [14]. In addition, other studies have accurately predicted apple pests and diseases [15,16], cotton pests and diseases by fusing convolutional neural network (CNN) and LSTM models, which demonstrated the ability of hybrid models to incorporate a variety of data sources to further improve prediction accuracy [17,18].

In addition, in predicting agricultural pests and diseases, not only the data of the crop itself, but also external environmental factors such as weather and climate change need to be considered. Studies have shown that weather data have a significant impact on the occurrence of pests and diseases [1,19] introduced the Standardized Precipitation Evapotranspiration Index (SPEI) through an ARIMA-LSTM model in a drought prediction study in China and demonstrated that this hybrid model can effectively predict the impact of climate change on crop growth [20,21]. The combination of ARIMA with models such as support vector regression (SVR) and least squares support vector machine (LS-SVR) has also demonstrated high prediction accuracy in studies on the effects of climatic factors on

pest and disease occurrence [10,22]. However, how to select appropriate models in specific agricultural scenarios and improve their prediction ability for complex pests and diseases remains an important challenge in current research [23].

This study presents an innovative hybrid ARIMA-LSTM model tailored for sugarcane pest and disease prediction, marking a significant advancement in agricultural forecasting. Unlike previous works that rely solely on either statistical or deep learning models, our approach systematically integrates the strengths of both methodologies. Moreover, we enhance prediction performance by incorporating exogenous climate variables, such as temperature and precipitation, into the modeling process. This hybrid framework not only improves forecasting accuracy but also provides actionable insights for precision agriculture, enabling farmers and policymakers to implement timely and effective disease control strategies. Our contribution lies in the novel fusion of time series statistical methods and deep learning techniques to develop a robust, data-driven decision-support tool for sustainable sugarcane production. In this study, we first modeled the linear characteristics of the pest and disease time series using the ARIMA model, and captured the potential nonlinear relationships and complex dynamics in the time series by LSTM [1,24]. In addition, we further enhanced the predictive power of the model by introducing exogenous variables such as weather data [15,19]. The approach in this study echoes previous studies, for example, the model combining ARIMA and dynamic support vector machine (SVM) performed well in predicting the tree pest *Dendrolimus punctatus* [25], while in predicting, for example, cotton pests and apple diseases, the LSTM hybrid model also showed highlight predictive ability [19,26]. Suresh and Priya predicted sugarcane yields in India through an ARIMA model, and although the results showed its effectiveness in capturing linear trends, the performance of a single model was still limited when faced with the complex prediction of pest and disease [3,27]. By combining ARIMA and LSTM, the hybrid model proposed in this paper is not only capable of capturing linear and nonlinear variations in pest and disease time series, but also incorporates external environmental factors to enhance the robustness of forecasts [23,28].

## 2. Materials and Methods

The prediction and management of sugarcane pests and diseases using big data represent a classic issue in time series data analysis, encompassing both linear and nonlinear elements. Previous research indicates that ARIMA serves as a conventional and potent linear statistical technique for forecasting time series. In contrast, the LSTM model is adept at identifying the nonlinear characteristics present in the data. Given that the emergence of pests and diseases in sugarcane is marked by intricate nonlinear variations, we suggest employing a hybrid ARIMA-LSTM model. This approach merges linear and nonlinear methodologies to enhance both the accuracy of pest and disease predictions and the efficiency of management strategies.

### 2.1. Data Collection and Preprocessing

To investigate the disease prediction for sugarcane using a hybrid ARIMA-LSTM model, this study collected time-series data on crop pests and disease incidences in relation to meteorological data containing four key variables: temperature, humidity, rainfall, and pest and disease incidences. The meteorological data were sourced from the National Meteorological Bureau's Agricultural Climate Monitoring Network, which provides daily observations recorded from a network of weather stations across major agricultural regions. Pest and disease incidence data were obtained through the China Agricultural Pest Forecasting and Control Network, part of the Ministry of Agriculture and Rural Affairs, which conducts systematic monitoring and publishes data on crop health and pest population

dynamics at a monthly and seasonal frequency. These data sources together simulate a long-term crop production environment to analyze the impact of meteorological conditions on crop pest and disease incidence. The entire dataset covers a period of approximately 33 years from 1 January 1990 to the present and contains a total of 16,000 records. From there, 75% of the data (12,000 records) is divided into a training set in chronological order and the remaining 25% of the data (4000 records) is used as a test set. This division ensures that the model is able to learn long-term series trends and thus make effective predictions on future data. Unlike conventional time-series data, the occurrence of crop pests and diseases is not only affected by meteorological factors, but may also be disturbed by other external factors, such as pest control measures, soil conditions, and crop species. However, in the present modeled data, these conditions are assumed to be relatively stable and fluctuations in pest and disease incidence are mainly driven by meteorological factors.

The time series of pest and disease incidents show significant fluctuations that are closely related to changes in weather conditions. For example, pest and disease incidence usually increases significantly under hot and humid climatic conditions. At the same time, increases or decreases in rainfall can also have an impact on crop health, which in turn can alter the rate of spread and severity of pests and diseases. Thus, to capture the linear and nonlinear relationships in pest and disease incidence, the study used a hybrid ARIMA-LSTM model. The ARIMA model was used to extract the linear trends in the meteorological data, while the nonlinear residual portion was fitted by the LSTM model. The objective of this hybrid model is to predict the future occurrence of pests and diseases, to help in the advancement of agricultural control measures, and to provide a basis for decision making for the healthy growth of crops.

The challenges posed by climate change and the expanding scope of global agriculture have made pests and diseases an increasing threat to agricultural productivity. Utilizing meteorological and crop growth data to forecast pest and disease occurrences has emerged as an effective strategy that enables farmers and agricultural managers to implement timely control measures, thereby mitigating large-scale outbreaks and minimizing crop losses.

When it comes to forecasting pests and diseases, the following variables play a vital role in the model's accuracy:

1. **Temperature:** This is a primary factor influencing the emergence of pests and diseases, with notable variations in their reproduction rates and dissemination speeds under various temperature conditions.
2. **Humidity:** In addition to affecting the cultivation environment of crops, humidity is also closely linked to the emergence and propagation of pests and diseases.
3. **Rainfall:** Both excessive and insufficient precipitation can have direct or indirect effects on the prevalence of pests and diseases.
4. **Pest Occurrence:** Information relating to the frequency and intensity of pest and disease events derived from meteorological data and crop health assessments.

These four variables were thus identified as the key elements for analysis. To ensure the consistency and precision of the predictive models, the gathered data underwent pre-processing, since both crop and meteorological information may contain missing values, anomalies, or temporal gaps. To address missing values, we applied linear interpolation, which estimates missing values based on the closest preceding and succeeding data points. For example, if the temperature data were missing for June 15, we computed its value as the average of June 14 and June 16. Additionally, to handle outliers, we used the  $3\sigma$  principle, where data points deviating beyond three standard deviations from the mean were flagged as anomalies. For instance, if a sudden temperature drop of  $15^{\circ}\text{C}$  was recorded within a single day, it was identified as an outlier and replaced using a rolling median

technique to maintain consistency. The specifics of linear interpolation are articulated by the following equation:

$$x_{\text{new}} = \frac{x_{\text{prev}} + x_{\text{next}}}{2} \quad (1)$$

where  $x_{\text{new}}$  was missing value and  $x_{\text{prev}}$  and  $x_{\text{next}}$  were the data points known before and after, respectively. Then, meteorological data may contain outliers, such as sudden increases or decreases in temperature. These outliers may reflect data acquisition errors or unusual climatic events. A commonly used method for detecting outliers is the  $3\sigma$  principle. When data exceeds  $\pm 3$  times, the standard deviation from the mean is considered as outliers. Thus, anomalous data can be effectively removed to ensure the robustness of the model. Finally, to adapt the data to the inputs of the model, it is necessary to normalize the range of values of the different variables. In this study, MinMaxScaler was used to normalize the data with the following equation:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

The normalized data are represented as  $x_{\text{norm}}$ , while  $x$  denotes the original data. The minimum and maximum values of the variable are indicated by  $x_{\min}$  and  $x_{\max}$ , respectively. Therefore, this normalization procedure enables comparison and modeling of the variables on a uniform scale.

## 2.2. ARIMA Model for Linear Component

The ARIMA model represents a traditional statistical approach used for analyzing time series data, making it especially effective for the modeling of linear time series. By incorporating both autoregressive (AR) and moving average (MA) components, the model captured the linear trends present in the data, while the integration (I) component was employed to address non-stationarity in time series. Through the ARIMA model, the linear aspect of the time series could be isolated, thereby simplifying the input for later nonlinear modeling tasks, such as those involving LSTM. The structure of the model was denoted as ARIMA ( $p, D, q$ ), with parameters  $p$ ,  $D$ , and  $q$  signifying the components of the predictive model that integrate autoregressive AR ( $p$ ), differencing  $D$ , and moving average MA ( $q$ ). The mathematical representation of this model was as follows:

$$\left( \sum_{i=1}^p \phi_i L^i \right) (1 - L)^D X_t = \left( 1 + \sum_{i=1}^q \theta_i L^i \right) \varepsilon_t \quad (3)$$

where  $L$  denotes the lag operator,  $\phi_i$  are the parameters of the autoregressive part of the model,  $\theta_i$  are the parameters of the MA part, and  $\varepsilon_t$  are error terms.

To construct an ARIMA model, the initial phase involves identifying and selecting the appropriate model type. Subsequently, the Augmented Dickey–Fuller (ADF) test is utilized to assess whether the time series exhibits stationarity. If the  $p$ -value obtained from the test is below the significance threshold (typically set at 0.05), the null hypothesis can be rejected, signifying that the series is stationary; conversely, if this is not the case, the series must undergo differentiation. The parameter indicates the number of differentiation operations required to achieve stationarity in the time series. In this context, we eliminate the trend by calculating the difference between values at consecutive time points, thereby stabilizing the series. Once the stationarity of the time series and the requisite degree of differencing  $d$  are established, it becomes necessary to select the autoregressive order  $p$  and the moving average order  $q$  for the ARIMA model. Various methodologies have been proposed, utilizing Akaike’s Information Criterion (AIC), Minimum Description Length (MDL), and Bayesian Information Criterion (BIC), among others. For this research, we

apply the AIC and BIC metrics to estimate the model parameters. After identifying these parameters, a diagnostic assessment is conducted through residual analysis to evaluate how well the model fits the data. Additionally, we assess the randomness of the residuals by either plotting their Autocorrelation Function (ACF) or executing a statistical test. Consequently, the ARIMA model is primarily employed to capture the linear dynamics of the time series. Following the model fitting, the residuals generated will be utilized as inputs for the subsequent LSTM model, addressing the nonlinear aspects of the time series.

### 2.3. LSTM Model for Nonlinear Component

The initial layer of the memory gate is responsible for filtering out unnecessary information from the cell state, which can be represented as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

where  $f_t$  represents the forgetting threshold at time  $t$ ,  $\sigma$  signifies the sigmoid activation function,  $W_f$  is the weight matrix for the gate,  $x_t$  indicates the incoming value,  $h_{t-1}$  refers to the output value at the previous time step, and  $b_f$  is the bias term.

The second input gate evaluates which information is to be retained in the cell state from the current input vector. This gate comprises two components: an input gate layer and a candidate value generation layer. The input gate layer assesses the fraction of the new input that will be incorporated into the unit state. This can be described as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)^2 \quad (5)$$

where  $i_t$  denotes the ratio of new information that is introduced. The values float between 0 and 1; a value of 0 signifies that no new information is provided, while a value of 1 indicates that entirely new information is introduced. The symbol  $\sigma$  represents the Sigmoid function, whose output values also range from 0 to 1, and it is employed to regulate the amount of new input information brought in. The term  $W_i$  refers to the weight matrix of the input gate, which is in charge of transforming both the current input and the prior hidden state into the necessary dimension. Meanwhile,  $h_{t-1}$  pertains to the hidden state from the preceding time step. The variable  $x_t$  represents the input for the ongoing time step, and  $b_i$  corresponds to the bias for the input. A layer for candidate values is utilized to formulate a new candidate value, which signifies the new information that may be included in the cell state. This can be articulated as follows:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (6)$$

where  $\tilde{C}_t$  is candidate cell state, which is a candidate value that the current time step can use to update the cell state. It is processed by the tanh function and has a value between  $-1$  and  $1$ , indicating the new candidate memory content.  $W_C$  describes the weight matrix of candidate unit states, responsible for transforming inputs and previous hidden states into candidate values. To update the state of the cell at time  $t$ , the expression is as follows:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (7)$$

where  $C_t$  represents the cell state at the present time step. It integrates the cell state from the previous time step, denoted as  $C_{t-1}$ , along with the newly created candidate value corresponding to the current time step. The value  $f_t$ , which originates from the forget gate, regulates how much of the previous cell state  $C_{t-1}$  is retained during the current time step. Meanwhile,  $i_t$  signifies the output from the input gate, which manages the extent to which

new information  $\tilde{C}_t$  is introduced. The term  $\tilde{C}_t$  reflects the status of the candidate cells that are generated during the current step.

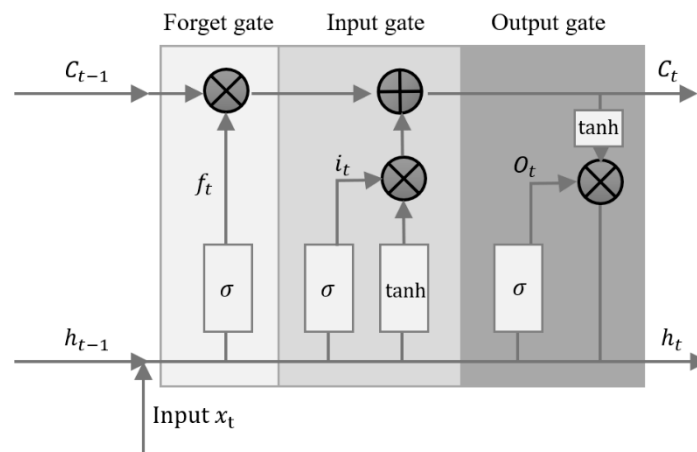
The output information for the third layer is generated in the present step and can be represented as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

where  $W_o$  is the weight matrix of the output gate and  $b_o$  is the bias of the output gate. Then, the updated hidden state can be described as follows:

$$h_t = o_t \cdot \tanh(C_t) \quad (9)$$

In this context,  $h_t$  represents the hidden state at the current timestep. The output value of the gate, denoted as  $o_t$ , regulates how much the hidden state will be updated. Meanwhile,  $C_t$  refers to the cell state at this particular moment, which retains the updated “memory.” Once the data has gone through the three gates, the resulting output contains pertinent information, while the irrelevant details are discarded (Figure 1).



**Figure 1.** Structure of an LSTM cell: representation of forget gate, input gate, and output gate.

#### 2.4. ARIMA-LSTM Model

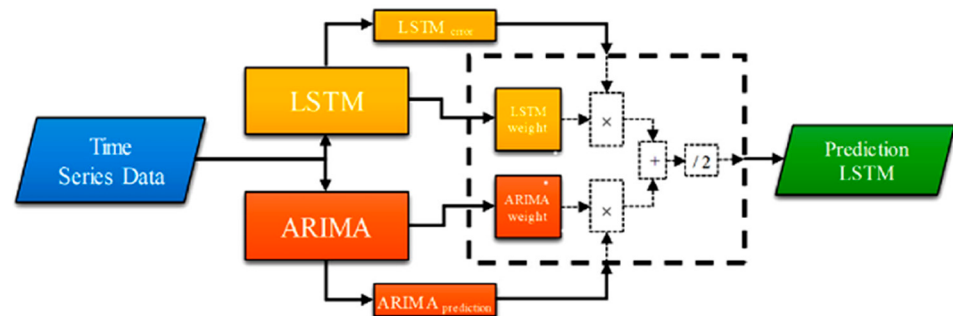
The ARIMA-LSTM model begins by employing the ARIMA component to identify the linear aspects of the time series. Subsequently, the leftover residuals serve as inputs that are processed by the LSTM model for nonlinear analysis. This approach allows LSTM to concentrate on the nonlinear components that the ARIMA model is unable to represent. Consequently, the ultimate predicted outcomes can be derived by combining the predictions from both the ARIMA model and the LSTM’s residual forecasts. The detailed formulation of the model can be outlined as follows:

$$\hat{y}_t = \hat{y}_{t, \text{ARIMA}} + \hat{y}_{t, \text{LSTM}} \quad (10)$$

where  $\hat{y}_{t, \text{ARIMA}}$  denotes the linearity of the data at time  $t$ ,  $\hat{y}_{t, \text{LSTM}}$  is nonlinearity.

To sum up, hybrid models can be categorized into several key phases (Figure 2). Initially, historical data are utilized to fit an ARIMA model. This model effectively captures linear trends along with seasonal and cyclical characteristics found in the data by employing techniques such as autoregression, differencing, and moving averages. After training the ARIMA model, it is applied to forecast the time series, producing predictions for its linear components. To enhance prediction accuracy, the differences between the actual values and the forecasts from the ARIMA model, known as residuals, are computed; these residuals indicate the nonlinear patterns that the ARIMA model does not account for. Subsequently,

an LSTM model is employed to analyze the residual series. The strength of LSTM lies in its ability to recognize long-term dependencies and intricate nonlinear relationships within time series data through its unique gating mechanism. During the training process, these residuals serve as inputs for the LSTM model to forecast future residual values. Ultimately, the linear forecasts derived from the ARIMA model are combined with the nonlinear predictions from the LSTM model to generate the final output of the hybrid model. This integration allows the hybrid approach to effectively capture linear trends while adeptly managing complex nonlinear behaviors, thereby enhancing prediction accuracy.



**Figure 2.** ARIMA-LSTM model.

### 2.5. Evaluation Indicators

To thoroughly assess the predictive efficacy of the hybrid model, we applied various evaluation metrics to the test set. These metrics illuminate the prediction errors and the model's performance from multiple angles. Frequently utilized assessment metrics encompass mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE), which serve to gauge the performance of diverse models in forecasting outcomes and can be articulated as follows:

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2 \quad (11)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (12)$$

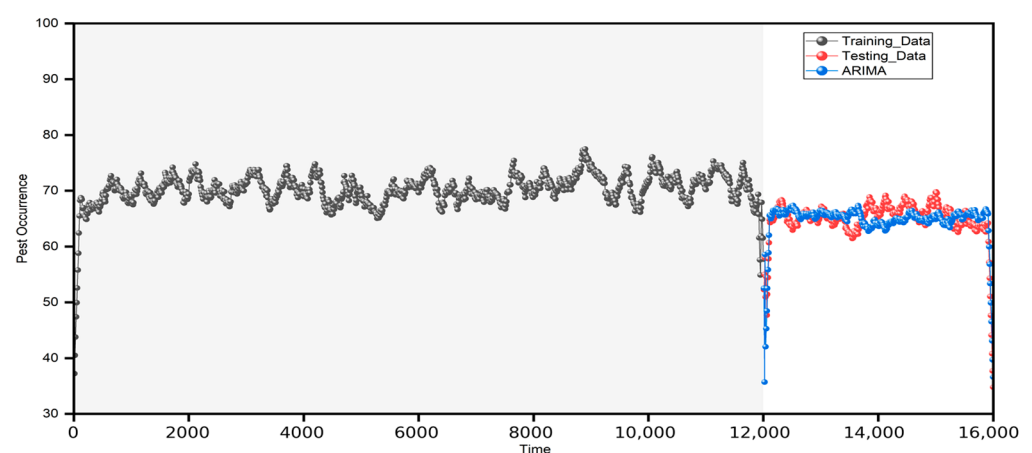
$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (13)$$

where  $y_t$  is actual observation at time  $t$  and  $\hat{y}_t$  is prediction value at time  $t$ .  $n$  is the amount of data. Typically, a lower mean squared error (MSE) indicates that the predictions made by a model are closer to the actual values; however, MSE is especially susceptible to significant errors, as it exaggerates larger discrepancies. In contrast, the root mean squared error (RMSE) shares the same measurement units as the original dataset, making it easier to interpret than the MSE. A reduced RMSE signifies greater accuracy in the model's predictions. In relation to MSE and RMSE, the mean absolute error (MAE) is often considered more robust since it does not magnify larger mistakes in the way MSE does. Consequently, MAE proves effective for evaluating a model's ability to consistently minimize prediction errors, particularly when only a few substantial errors are present. By leveraging a combination of these evaluation metrics, one can attain a more thorough understanding of a model's predictive capabilities, examining its various strengths and weaknesses from different angles.

### 3. Results

#### 3.1. Linear Prediction of ARIMA Model

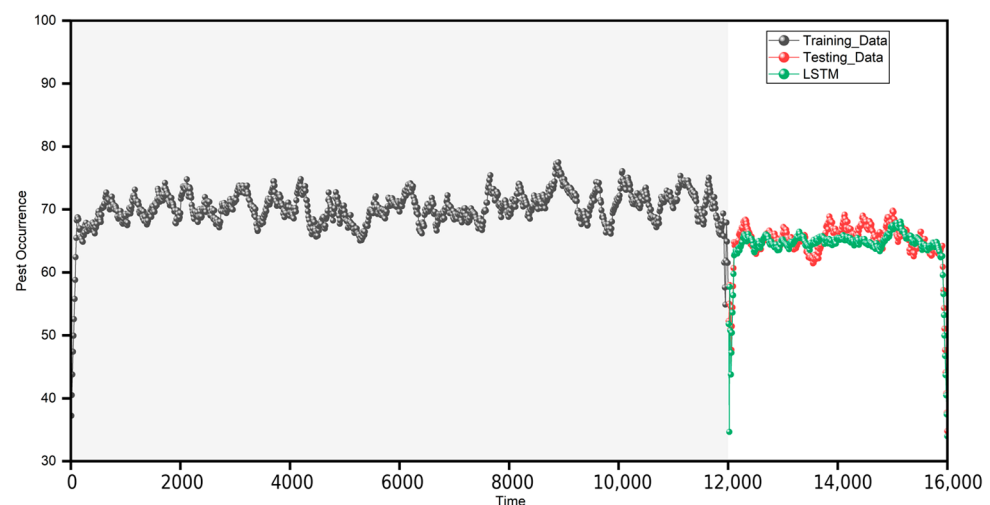
The ARIMA model was utilized to derive the linear predicted values. The results of predicting pest and disease incidence through the ARIMA model are depicted in Figure 3. This model encompasses three primary stages: identification, estimation, and prediction. Initially, a distinct operation is employed to ensure the time series remains smooth and to ascertain the order of integration, denoted as  $D$ . Commonly,  $D$  can take the values of 0, 1, or 2 to confirm that the data satisfy smoothness requirements. In this illustration, the training dataset underwent differentiation and was subsequently fitted using the ARIMA approach to generate linear predicted values and residuals; these outcomes were then used for predicting the test data. The findings indicate that the ARIMA model is predominantly effective in addressing linear trends within time series data. In the figure, the training dataset is represented by the black dotted line, the test dataset by the red dotted line, and the prediction results produced by the ARIMA model are shown as a blue curve. The ARIMA model substantially outperforms the training dataset and provides a more precise prediction of the test dataset's linear trend. The model demonstrates a strong alignment with the test data after 12,000 days, particularly in the initial stages when it almost flawlessly predicts the trend of the test data. However, as time passes, the model is slightly insufficient in handling certain nonlinear fluctuations, leading to some deviations between the predicted results and the actual data at the later stage. The error analysis further demonstrates the performance of the ARIMA model. The MSE is 4.97, indicating that the overall prediction error of the model is small; the RMSE is 2.29, indicating that the standard deviation of the predicted values is small. The prediction results have a small deviation from the actual data. And the MAE is 1.79, indicating that the average prediction error is relatively low. In practice, the ARIMA model has a good prediction effect on smooth linear trends in time series, but its performance is relatively limited for complex nonlinear fluctuations. Therefore, the ARIMA model is suitable for relatively smooth time series but may need to be supplemented and optimized by introducing other models when more complex nonlinear variations need to be dealt with. Overall, the ARIMA model provides more reliable linear forecasts, but to further improve the performance, it is recommended to incorporate other nonlinear models to better capture the complex fluctuations in the time series.



**Figure 3.** ARIMA model result.

### 3.2. Nonlinear Prediction of LSTM Model

Figure 4 illustrates the prediction outcomes related to pest and disease occurrences generated by the LSTM model, a deep learning framework designed for time series analysis that excels in capturing both short- and long-term dependencies while effectively managing nonlinear characteristics. This figure indicates that both the training data (represented by the black dotted line) and the testing data (depicted by the red dotted line) tend to display a steady performance throughout most of the duration, with minimal fluctuations. The LSTM model's predictions are represented by the green dotted lines, spanning the complete duration of the test data. From a visual standpoint, the LSTM model demonstrates a high degree of accuracy in predicting the test data, particularly during the initial prediction phase, where it effectively tracks the variations in the actual data. Nevertheless, as time progresses, some discrepancies in predictions appear in later stages, especially at points of greater fluctuation, where the LSTM's forecasted values exhibit slight deviations from the observed values. Regarding error analysis, the MSE (mean square error) registered for the LSTM model is 3.77, the RMSE (root mean square error) stands at 1.86, and the MAE (mean absolute error) is 1.45. These figures suggest that the LSTM model surpasses the ARIMA model in handling intricate nonlinear relationships. Specifically, the comparatively low MSE and RMSE values demonstrate that the LSTM model can more accurately predict pest and disease occurrences across the majority of time intervals. Moreover, an MAE of 1.45 signifies that the average prediction error is below 2, which is highly favorable for time-series forecasting.



**Figure 4.** LSTM model result.

In general, the performance of the LSTM model was commendable when handling the nonlinear characteristics of the time series, particularly in its ability to detect short-term variations and long-term patterns within the dataset. Nevertheless, the model's forecasts exhibit some bias during periods of extreme volatility, indicating that additional optimization of its parameters or adjustments to the training data could be necessary to enhance prediction accuracy further. In contrast to the ARIMA model, LSTM proves to be more effective for analyzing time series data that exhibit intricate fluctuations, particularly in situations where there are nonlinear relationships to be identified.

### 3.3. Prediction Results of ARIMA-LSTM Model

Figure 5 illustrates the outcomes of predicting pest and disease occurrences using the hybrid ARIMA-LSTM model. This model integrates the ARIMA's capacity to capture linear elements with the LSTM's strength in handling nonlinear dynamics, thus enabling it to effectively manage complex time series data. The black dotted line in the figure indicates the training dataset, while the red dotted line signifies the testing dataset, and the purple dotted line represents the predictions made by the ARIMA-LSTM model. It is evident from the visual data that the ARIMA-LSTM model surpasses the performance of the individual ARIMA or LSTM models. Initially, the model adeptly identifies both the overall trajectory and the localized variations of the test data after 12,000 days; the prediction curve aligns closely with the actual test data. In the preliminary phase of forecasting, the purple prediction curve nearly coincided with the test data, suggesting the model's competence in accurately forecasting the trends and variations of pest and disease incidents, particularly during periods of relative stability, with minimal prediction error. Moreover, even as data fluctuations increased in the later stages, the ARIMA-LSTM model continued to exhibit high accuracy, demonstrating its adaptability to both steady and highly dynamic periods. Unlike ARIMA, which struggles with sharp variations, and LSTM, which occasionally fails to capture long-term trends, the hybrid approach effectively balances both aspects, ensuring improved generalization across different time periods. The analysis of errors further highlights the advantages of the ARIMA-LSTM approach. The model's mean square error (MSE) is recorded at 2.66. The root mean square error (RMSE) stands at 1.63, and the mean absolute error (MAE) is 1.34. These performance metrics are more favorable compared to those of the ARIMA model, which has an MSE of 4.97, an RMSE of 2.29, and an MAE of 1.79, as well as the LSTM model, which presents an MSE of 3.77, an RMSE of 1.86, and an MAE of 1.79. The reduction in MSE suggests that the ARIMA-LSTM model demonstrates superior overall prediction accuracy and a reduced gap between the expected and actual values. In addition, the lower RMSE and MAE signify a decrease in error variability, with the average prediction error remaining consistently low across all time intervals. Further investigation reveals that the strength of the ARIMA-LSTM model is its capacity to manage both linear and nonlinear characteristics within time series data. The ARIMA model is proficient at addressing the linear aspects of the data, particularly in identifying trends and seasonal variations; nevertheless, it faces challenges when dealing with intricate nonlinear fluctuations. Conversely, the LSTM model, with its recursive neural network architecture, retains long-term dependencies within the time series and adeptly manages nonlinear data features. By integrating ARIMA and LSTM, the ARIMA-LSTM model effectively addresses both linear and nonlinear aspects, significantly enhancing the model's robustness and forecasting precision.

Moreover, an additional benefit of the hybrid model is its robust generalization capability. The prediction curves depicted in the figure not only provide an accurate representation of the training data but also enhance the prediction of future data fluctuations. This suggests that the ARIMA-LSTM model is not solely reliant on the training dataset for modeling but is also adept at successfully predicting new data with consistent performance. This capability is particularly important in the context of real-world pest and disease prediction challenges, as the emergence of crop pests and diseases is influenced not only by linear climate patterns but also by intricate factors such as climate anomalies and seasonal variations. Consequently, the versatile adaptive capacity of ARIMA-LSTM models enables their broad application in time series forecasting for pest and disease management.

In summary, the hybrid ARIMA-LSTM framework considerably enhances the precision and durability of time series forecasts by integrating the strengths of both ARIMA and LSTM models. This model is proficient in addressing linear trends in the data while

simultaneously managing nonlinear fluctuations, thereby ensuring that it delivers more precise forecasts across various time periods.

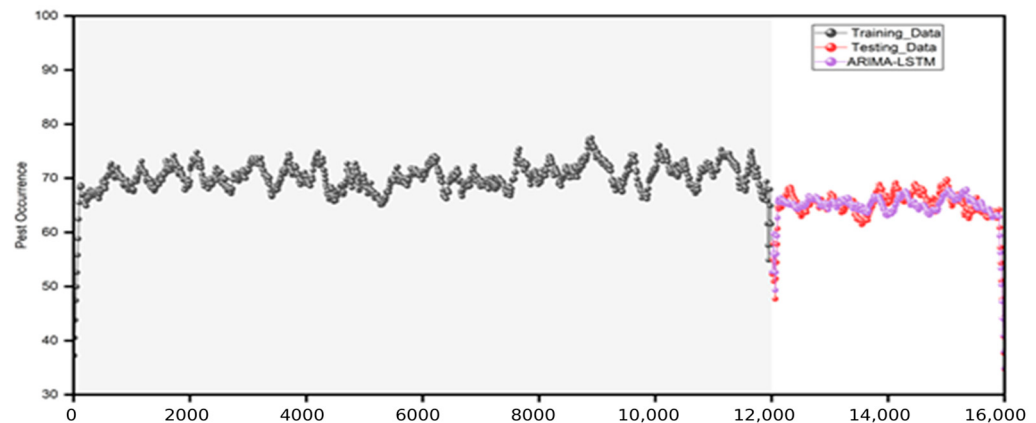


Figure 5. ARIMA-LSTM model result.

#### 4. Discussion

In this study, the performance of ARIMA model, LSTM model and ARIMA-LSTM hybrid model in the prediction of time series data, especially in the prediction of sugarcane pests and diseases, was comparatively analyzed. The ARIMA model is a traditional time series forecasting method that deals mainly with linear trends. It performs well in capturing long-term trends and seasonal fluctuations, especially with data smoothing. However, the limitation of the ARIMA model is its inability to handle complex nonlinear fluctuations. This limitation is reflected in its high error values of 4.97 for MSE, 2.29 for RMSE, and 1.79 for MAE, which indicates its poor prediction accuracy when dealing with nonlinear fluctuations in pests and diseases. In contrast, the LSTM model, as a type of recurrent neural network, excels in managing nonlinear dependencies and capturing short-term fluctuations through its ‘gating mechanism’. However, when dealing with extended sequences, the LSTM model can sometimes lose precision in long-term trend estimation, leading to deviations from actual values. This issue is particularly noticeable when the dataset exhibits strong seasonal patterns, where ARIMA generally performs better. The hybrid model overcomes this shortcoming by allowing ARIMA to handle the trend component while LSTM focuses on residual nonlinearity, ensuring more comprehensive predictions. The model has a low error value of 3.77 MSE, 1.86 RMSE and 1.45 MAE, which shows its superiority in dealing with complex nonlinear relationships. However, the LSTM model may exhibit biases in scenarios where linear trends dominate over a prolonged period, as it primarily focuses on detecting and responding to short-term variations. Additionally, the LSTM model requires significantly more training data and computational resources compared to ARIMA, which makes it less efficient for datasets with limited historical records. The ARIMA model, while computationally efficient, fails to handle complex nonlinearities, making it inadequate for predicting abrupt fluctuations. By merging the capabilities of both models, the hybrid ARIMA-LSTM framework successfully achieves a balance between computational efficiency and predictive power, minimizing errors across varying trend complexities. The hybrid ARIMA-LSTM model combines the linear trend capturing ability of ARIMA and the advantage of LSTM in dealing with nonlinear relationships, and thus its prediction accuracy is significantly improved. In the error analysis, the MSE of the hybrid model is 2.66, the RMSE is 1.63, and the MAE is 1.34, which is significantly better than the results of using ARIMA or LSTM models alone (Table 1). This indicates that the hybrid model is not only able to effectively deal with the linear features in the data, but also able to cope with the complex nonlinear fluctuations. In addition, the hybrid model demonstrates

strong generalization ability and is able to maintain stable prediction performance on new datasets.

**Table 1.** Terrors of prediction results using three different models.

| Model      | MSE  | RMSE | MAE  |
|------------|------|------|------|
| ARIMA      | 4.97 | 2.29 | 1.79 |
| LSTM       | 3.77 | 1.86 | 1.45 |
| ARIMA-LSTM | 2.66 | 1.63 | 1.34 |

While the ARIMA-LSTM hybrid model significantly outperforms the standalone ARIMA and LSTM models, it is not without limitations. One notable issue observed is the tendency to underestimate pest and disease occurrences during periods of high volatility. This is likely due to the ARIMA component's reliance on past linear trends, which may not adequately capture abrupt shifts, and the LSTM component's potential over-smoothing in extreme fluctuation scenarios. These underestimations could impact decision-making in agricultural management, particularly in scenarios requiring immediate intervention. Future improvements could include the incorporation of real-time anomaly detection mechanisms or adaptive weighting between ARIMA and LSTM contributions to dynamically adjust to sudden fluctuations.

The ARIMA-LSTM hybrid model has several advantages over individual models, notably its ability to capture both linear trends and complex nonlinear fluctuations, resulting in lower prediction errors. However, this comes at the cost of increased computational requirements. The LSTM component, in particular, demands significant computational power and extensive training data, making real-time implementation in resource-constrained environments challenging. Additionally, while the model is tailored for sugarcane pest and disease prediction, its scalability to other crops requires careful consideration. Variability in pest dynamics, environmental interactions, and crop-specific growth patterns may necessitate additional model tuning and the inclusion of crop-specific parameters. Future research should explore strategies for optimizing computational efficiency, such as model pruning or using lighter deep learning architectures without compromising prediction accuracy.

The findings of this study have direct implications for agricultural practices and policymaking. By providing more accurate pest and disease forecasts, the hybrid model enables farmers to implement preventive measures, optimize pesticide usage, and mitigate crop losses. This predictive capability supports precision agriculture, reducing unnecessary pesticide applications, thus lowering costs and minimizing environmental impact. From a policy perspective, integrating such predictive models into national agricultural monitoring systems could enhance early warning systems and disaster preparedness for pest outbreaks. Additionally, policymakers could use these insights to develop targeted subsidies or support programs for farmers facing high-risk pest infestation periods. The hybrid model thus not only benefits individual farmers but also contributes to broader agricultural sustainability and food security strategies. Future work should focus on enhancing the robustness of the ARIMA-LSTM hybrid model by incorporating additional agronomic and environmental variables. Soil quality, fertilizer applications, and crop variety data could further refine predictive accuracy by accounting for factors that influence pest and disease susceptibility. Additionally, integrating remote sensing and IoT-based real-time monitoring data could enhance the model's responsiveness to environmental changes. Another promising direction is the exploration of ensemble learning, combining ARIMA-LSTM with other machine learning approaches, such as gradient boosting or transformers, to further improve predictive power. These advancements will ensure the model remains adaptable across different agricultural contexts, reinforcing its value in precision agriculture.

In conclusion, the use of hybrid ARIMA-LSTM models for predicting agricultural pests and diseases holds significant potential. As big data and deep learning technologies continue to advance, these hybrid models can further fine-tune parameter settings and enhance computational capabilities, particularly in the context of complex climate shifts and extreme weather phenomena. Moreover, integrating additional environmental factors, like soil conditions and the application of fertilizers, is anticipated to enhance the models' predictive accuracy. Given the escalating challenges posed to agricultural output by global climate change, such hybrid approaches will offer improved accuracy in predictions and management strategies for agricultural practices, enabling farmers to implement effective preventive measures and control strategies proactively to mitigate losses due to pests and diseases.

## 5. Conclusions

This research highlighted the advantages of hybrid models through an evaluation of the effectiveness of ARIMA, LSTM, and hybrid ARIMA-LSTM models in predicting sugarcane pests and diseases. The ARIMA approach effectively manages linear trends within time series but struggles with complex nonlinear variations. In contrast, the LSTM model excels at identifying nonlinear characteristics yet faces challenges in forecasting long-term linear trends. By merging the strengths of both methodologies, the hybrid ARIMA-LSTM model considerably enhances prediction accuracy, achieving lower error metrics (MSE, RMSE, and MAE) compared to singular models, thus showcasing its capability to address both linear and nonlinear features effectively. This model not only adjusts to various types of time series data but also exhibits a robust generalization capacity, ensuring consistent prediction performance even with new data inputs. With ongoing enhancements in data dimensions and algorithm optimization, this model holds significant potential for applications in the prediction and management of agricultural pests and diseases, offering more precise decision-making support for farmers and agricultural managers while effectively tackling the challenges posed by global climate change.

**Author Contributions:** M.W.: Data curation, Investigation, Methodology, Visualization, Writing—original draft, Writing—review and editing. T.L.: Funding acquisition, Software, Validation, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Special Project of the Key Laboratory for Crop Production and Smart Agriculture of Yunnan Province (202105AG070007) -the Central Guided Local Science and Technology Development Fund “Key Technologies of Citrus Big Language Modeling and Service System” (202407AB110010) -the Major Scientific and Technological Project in Yunnan Province “Key Technology Integration of Intelligent Agriculture” (202302AE090020).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original data presented in the study are openly available in the National Meteorological Bureau's Agricultural Climate Monitoring Network and the China Agricultural Pest Forecasting and Control Network, part of the Ministry of Agriculture and Rural Affairs. Access to meteorological data can be found at [<http://www.fema.gov/>], accessed on 2 July 2024], and pest and disease incidence data are available at [<https://www.technavio.com/report/agricultural-pesticides-market-industry-analysis>], accessed on 2 July 2024].

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Xu, D.; Zhang, Q.; Ding, Y.; Zhang, D. Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environ. Sci. Pollut. Res.* **2021**, *29*, 4128–4144. [CrossRef] [PubMed]
- Donatelli, M.; Magarey, R.D.; Bregaglio, S.; Willocquet, L.; Whish, J.; Savary, S. Modelling the impacts of pests and diseases on agricultural systems. *Agric. Syst.* **2017**, *155*, 213–224. [CrossRef]
- Suresh, K.K.; Priya, S.R.K. Forecasting sugarcane yield of Tamilnadu using ARIMA models. *Sugar Tech* **2011**, *13*, 23–26. [CrossRef]
- Materne, N.; Inoue, M. Potential of IoT System and Cloud Services for Predicting Agricultural Pests and Diseases. In Proceedings of the IEEE Region Ten Symposium (Tensymp), Sydney, NSW, Australia, 4–6 July 2018; pp. 298–299.
- Fan, D.; Sun, H.; Yao, J.; Zhang, K.; Yan, X.; Sun, Z.X. Well production forecasting based on ARIMA-LSTM model considering manual operations. *Energy* **2021**, *220*, 119708. [CrossRef]
- Grünig, M.; Razavi, E.; Calanca, P.; Mazzi, D.; Wegner, J.D.; Pellissier, L. Applying deep neural networks to predict incidence and phenology of plant pests and diseases. *Ecosphere* **2021**, *12*, e03791. [CrossRef]
- Fenu, G.; Mallocci, F. An Application of Machine Learning Technique in Forecasting Crop Disease. In Proceedings of the 3rd International Conference on Big Data Research, Paris, France, 20–22 November 2019.
- Temraz, M.; Kenny, E.; Ruelle, E.; Shalloo, L.; Smyth, B.; Keane, M. Handling Climate Change Using Counterfactuals: Using Counterfactuals in Data Augmentation to Predict Crop Growth in an Uncertain Climate Future. In *Case-Based Reasoning Research and Development: 29th International Conference, ICCBR 2021, Salamanca, Spain, 13–16 September 2021*; Proceedings 29; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 216–231.
- Jain, S.; Ramesh, D. AI based hybrid CNN-LSTM model for crop disease prediction: An ML advent for rice crop. In Proceedings of the 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 6–8 July 2021.
- Jin, Y.; Wang, R.; Zhuang, X.; Wang, K.; Wang, C. Prediction of COVID-19 data using an ARIMA-LSTM hybrid forecast model. *Mathematics* **2022**, *10*, 4001. [CrossRef]
- Sara, D.; Maharani Mdd Amin, H.F.; Triana, Y.S. Application of Artificial Intelligence in Modern Ecology for Detecting Plant Pests and Animal Diseases. *Int. J. Quant. Res. Model.* **2021**, *2*, 83–90. [CrossRef]
- Garrett, K.; Bebbler, D.; Etherton, B.; Gold, K.; Sulá, A.I.P.; Selvaraj, M. Climate Change Effects on Pathogen Emergence: Artificial Intelligence to Translate Big Data for Mitigation. *Annu. Rev. Phytopathol.* **2022**, *60*, 357–378. [CrossRef]
- Guo, Y.; Feng, Y.; Qu, F.; Zhang, L.; Yan, B.; Lv, J. Prediction of hepatitis E using machine learning models. *PLoS ONE* **2020**, *15*, e0237750. [CrossRef]
- Dhawan, N.; Kukreja, V.; Sharma, R.; Vats, S.; Verma, A. Deep learning-based sugarcane downy mildew disease detection using CNN-LSTM ensemble model for severity level classification. In Proceedings of the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023.
- Taylor, R.A.; Ryan, S.; Lippi, C.; Hall, D.; Narouei-Khandan, H.A.; Rohr, J.R.; Johnson, L. Predicting the fundamental thermal niche of crop pests and diseases in a changing world: A case study on citrus greening. *J. Appl. Ecol.* **2019**, *56*, 2057–2068. [CrossRef]
- Peters, D.; McVey, D.S.; Elias, E.; Pelzel-McCluskey, A.; Derner, J.; Burruss, N.; Schrader, T.S.; Yao, J.; Pauszek, S.; Lombard, J.; et al. Big data–model integration and AI for vector-borne disease prediction. *Ecosphere* **2020**, *11*, e03157. [CrossRef]
- Skendžić, S.; Zovko, M.; Živković, I.; Lešić, V.; Lemić, D. The Impact of Climate Change on Agricultural Insect Pests. *Insects* **2021**, *12*, 440. [CrossRef]
- Hodson, D.; White, J.W.; Reynolds, M. GIS and crop simulation modelling applications in climate change research. *Clim. Change Crop Prod.* **2010**, 245–262. Available online: <https://www.cabidigitallibrary.org/doi/abs/10.1079/9781845936334.0245> (accessed on 12 February 2025).
- Chen, P.; Xiao, Q.; Zhang, J.; Xie, C.; Wang, B. Occurrence prediction of cotton pests and diseases by bidirectional long short-term memory networks with climate and atmosphere circulation. *Comput. Electron. Agric.* **2020**, *176*, 105612. [CrossRef]
- Tonnang, H.; Hervé, B.; Biber-Freudenberger, L.; Salifu, D.; Subramanian, S.; Ngowi, V.; Guimapi, R.; Anani, B.; Kakmeni, F.; Affognon, H.; et al. Advances in crop insect modelling methods—Towards a whole system approach. *Ecol. Model.* **2017**, *354*, 88–103. [CrossRef]
- Jung, J.; Maeda, M.; Chang, A.; Bhandari, M.; Ashapure, A.; Landivar-Bowles, J. The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Curr. Opin. Biotechnol.* **2020**, *70*, 15–22. [CrossRef]
- Bukhari, A.H.; Raja, M.; Sulaiman, M.; Islam, S.; Shoaib, M. Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access* **2020**, *8*, 71326–71338. [CrossRef]
- Sakshi, K.; Vijayalakshmi, A. An ARIMA-LSTM hybrid model for stock market prediction using live data. *J. Eng. Sci. Technol. Rev.* **2020**, *13*, 61–64. [CrossRef]
- Vivek, S.; Jesma, V. The changing phase of agriculture: Artificial intelligence. *AgricINTERNATIONAL* **2019**, *6*, 33. [CrossRef]
- Xiang, C.S.; Zhou, Z.Y.; Wu, L.N. Dendrolimus punctatus forecasting based on hybrid ARIMA and dynamic SVM model. *J. Hunan Agric. Univ.* **2010**, *36*, 430–433.

26. Turkoglu, M.; Hanbay, D.; Şengur, A. Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests. *J. Ambient Intell. Humaniz. Comput.* **2019**, *13*, 3335–3345. [[CrossRef](#)]
27. Scherm, H. Climate change: Can we predict the impacts on plant pathology and pest management? *Can. J. Plant Pathol.* **2004**, *26*, 267–273. [[CrossRef](#)]
28. Rani, R.; Sahoo, J.; Bellamkonda, S.; Kumar, S.; Pippal, S. Role of Artificial Intelligence in Agriculture: An Analysis and Advancements With Focus on Plant Diseases. *IEEE Access* **2023**, *11*, 137999–138019. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.