Research on World Agricultural Economy
https://journals.nasspublishing.com/index.php/rwae

RESEARCH ARTICLE

Time Series Analysis and Optimization of the Prediction Model of Agricultural Insurance Loss Ratio

Yu Wang, Muhammad Asraf Bin Abdullah* , *Josephine Yau Tan Hwang

Faculty of Economics and Business, Universiti Malaysia Sarawak (UNIMAS), Kota Samarahan, Sarawak 94300, Malaysia

ABSTRACT

For ensuring successful financial planning to perform sustainable farming, one key sector is to provide solutions that could accurately predict the agricultural loss ratios. In China, the Henan province is considered to be an agricultural center that is primarily exposed to drastic weather fluctuations that directly impact the crop yields. This study was conducted in Henan province from January 2020 to December 2023. With the data collected from that period, the study proposes a combinatory model combining Deep Gaussian Processes with Bayesian Long Short-Term Memory (LSTM) networks. The model was trained on data encompassing weather conditions, agricultural practices, and historical insurance claims. The experimental analysis was conducted against other traditional models, including ARIMA and Support Vector Regression. The RMSE improvement of the proposed model was around 7.2% on training data and 8.2% on test data, which demonstrates enhanced predictive accuracy. The enhanced performance of the proposed model was reflected in its effectiveness in reducing log-likelihood errors across training epochs. The model had demonstrated better robustness in handling complex and multi-dimensional agricultural data.

Keywords: Agriculture; Bayesian LSTM; ARIMA; Support Vector Regression; Loss Ratio; Log-Likelihood Errors

*CORRESPONDING AUTHOR:
Muhammad Asraf Bin Abdullah, Faculty of Economics and Business, Universiti Malaysia Sarawak (UNIMAS), Kota Samarahan, Sarawak 94300, Malaysia; Email: amasraf2024@outlook.com

ARTICLE INFO
Received: 6 August 2024 | Revised: 5 September 2024 | Accepted: 6 September 2024 | Published Online: 7 November 2024
DOI: https://doi.org/10.36956/rwae.v5i4.1219

CITATION
Wang, Y., Abdullah, M.A.B., Hwang, J.Y.T., 2024. Time Series Analysis and Optimization of the Prediction Model of Agricultural Insurance Loss Ratio. Research on World Agricultural Economy. 5(4): 299–312. DOI: https://doi.org/10.36956/rwae.v5i4.1219

1. Introduction

In the domain of agricultural economics, insurance schemes seem to have contributed to mitigating farmers' losses due to many unforeseen circumstances and market volatility [1]. In Asia, China has for many years projected itself as an agricultural hub, and regions such as Henan Province have contributed a lot to the nation's economy through agriculture. The influence of unpredictable weather patterns and fluctuating market conditions makes agricultural insurance models a must-have aspect of modern agriculture practices. The Henan province is a region that contributes higher wheat production, which is highly susceptible to losses due to events such as droughts and floods, which can devastate crops and, by extension, farmers' livelihoods. Under such circumstances, the only tool of savior is considered to be the insurance schemes.

Recent studies in agricultural insurance have adopted varied methodologies to tackle industry challenges. King and Singh [2] have explored how the behavioral factors have influenced the feature of demand for agricultural index insurance by understanding the roles of private transfers and farmer union memberships. In another study of Zhong et al. [3], they have examined how the government insurance subsidy schemes should be structured. They had found that such subsidy schemes are often misaligning with interests of farmers who are mainly worried about varying yield uncertainties and weather hazards. The work by Möhring et al. [4] have identified a positive link between the crop insurance enrollment and increased pesticide use in Europe; this link suggests that the insurance may inadvertently raise pesticide costs by up to 11%. The study of Chowdhury, Mayilvahanan and Govindaraj [5] has involved enhancing the health insurance predictions model by using Internet of Things (IoT) and Machine Learning (ML) models in the process of optimizing the Feature Extraction (FE) methodology employing a Whale Optimization Algorithm (WOA) for better accuracy. Dhieb et al. [6] have introduced a blockchain-based insurance system that uses XGBoost for detecting fraudulent claims; this detecting mechanism has enhanced the security and efficiency.

Also, advancements in the field of data analytics have driven the usage of tools like statistical and ML models for the process of predicting the insurance Loss Ratios (LR) with better accuracy and efficiency [7–9]. But still, many models have shown to have limitations in handling the dynamic and complex nature of agricultural data, which includes temporal sequences and multi-dimensional variables that are prejudiced by human activities and environmental factors [10, 11]. Another primary alarm that makes the existing models struggle is those related to the high variability and uncertainty inherent in agricultural processes because such uncertainties result in poor prediction against real-world outcomes. This limitation had led to the motivation of this study, which involved employing a hybrid analytical framework that combines Bayesian Long Short-Term Memory (LSTM) networks with Deep Gaussian Processes (DGP). The model takes the advantage of LSTM's sequential data processing and Gaussian Processes probabilistic modeling capabilities.

The study was carried out during a period from January 2020 to December 2023 around Henan Province, China's top wheat-producing region [11]. The model builds a dataset that includes variables such as weather conditions, farming practices, and historical LR. As influenced by the motivation to handle variability and uncertainty impacted by climatic and other circumstantial events, the work proposed a novel predictive model: a DGP integrated with a Bayesian LSTM network [12]. The model handles the complexities and uncertainties inherent in the temporal and multi-dimensional agricultural data. The proposed model was assessed for its prediction accuracy in insurance LR, which showed that it had significant prediction ability compared to other statistical and ML models [13–15].

The Objectives are:

(a) **Cost-Effectiveness Analysis:** Using statistical analysis, compare different interventions or strategies in research to determine the most cost-effective approach and highlight the most economic value based on the findings.

(b) **Economic Forecasting:** Utilize statistical models like time series analysis to predict future market conditions based on historical data, and discuss the implications for stakeholders.

(c) **Impact Assessment:** Evaluate the economic im-

pact of changes in variables like climate or crop varieties, using statistical methods to quantify potential financial impacts on producers, consumers and economy.

The work is projected in the following sections: Section 2 presents the theoretical framework; Section 3 presents the methodology; Section 4 presents the study's analysis and findings; Section 5 concludes the article.

2. Methods

2.1. LSTM Model

LSTMs have a chain-like structure (**Figure 1**) but replace the traditional nodes of RNNs with memory cells. These cells can maintain information in memory for long periods. Each cell has three types of gates: Input Gate (IG), Output Gate (OG), and Forget Gate (FG). These gates determine whether or not to let new input in (IG), delete the information because it is no longer necessary (FG), or let it impact the output at the current timestep (OG).

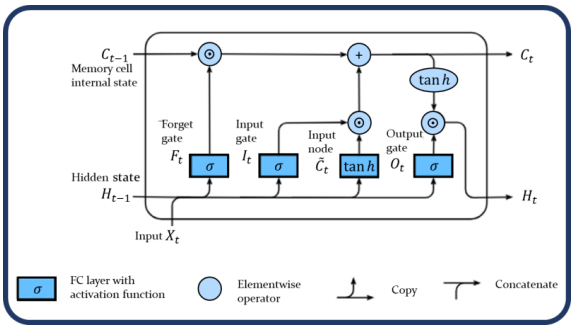

Figure 1. LSTM architecture.

1 FG (f_t): The FG is the first critical component, deciding which information the cell should discard. Using the sigmoid function 'σ', it looks at the previous hidden state $_{t-1}$ and the current input 'x_t'and outputs a number between 0 and 1 for each number in the cell state C_{t-1}, Equation (1).

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

A value close to 1 means "keep this completely", while a value close to 0 means "completely forget this". This decision-making process allows the LSTM to dynamically reduce the influence of less relevant past information.

2 IG (i_t) and Candidate Cell State $\widetilde{C}_t$: Simultaneously, the IG decides which new information is stored in the cell state. It operates similarly to the FG but determines where to add information, Equation (2).

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2}$$

Alongside, a candidate cell state $\widetilde{C}_t$ is created by a *tanh* layer, proposing a vector of new values to be added to the state, Equation (3).

$$\widetilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{3}$$

This candidate blends the old state and new insights from current inputs, prepared to update the cell state.

3 Cell State Update C_t : The cell state C_t is the next to be updated, integrating decisions from both the FG and IG, Equation (4).

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \tag{4}$$

This strategically updates FG parts of the previous state while adding new relevant information, ensuring the cell state carries forward only the necessary data.

4 OG (o_t) and Hidden State h_t : Finally, the OG determines what part of the current cell state to output as the hidden state h_t, Equation (5) and Equation (6).

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{5}$$

$$h_t = o_t \cdot tanh\left(C_t\right) \tag{6}$$

This step filters the cell state through another tanh function to regulate the network's output, allowing the LSTM to control the influence of its internal state on the outputs and subsequent states (**Figure 2**).

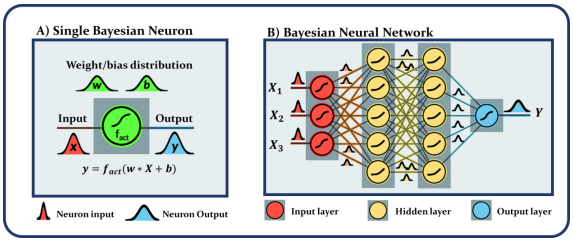

Figure 2. Bayesian neural network.

2.2. Bayesian Neural Network (BNN)

A BNN introduces probabilistic inference by treating the model weights as random variables with prior distributions (**Figure 2**). This helps the BNN model to achieve better handling of uncertainty and overfitting as against other traditional Neural Networks (NN) [16–18]. The foundation of Bayesian inference in NN starts with a prior distribution on the weights W, typically assumed to be Gaussian, Equation (7).

$$W \sim N\left(0, \sigma^2 I\right) \tag{7}$$

This assumption means that, before seeing any data, the weights are predicted to vary around zero with a variance 'σ^2', encapsulated within an identity matrix 'I'. This prior reflects our initial beliefs about the weights' distribution before data observation [19–26]. The likelihood function quantifies how probable observed data 'D' is, given a set of weights 'W'. The NN models this relationship directly through its architecture and activation functions, Equation (8).

$$y = f(x, W) + \epsilon \tag{8}$$

Here, 'ϵ' represents the noise associated with the outputs, assumed to be Gaussian, $N\left(0, \sigma_y^2\right)$. This model encapsulates the assumption that the true relationship is realized through the neural network function 'f', perturbed by some random Gaussian noise. Integrating the prior distribution and the likelihood of the observed data, the posterior distribution of the weights is updated using Bayes' theorem, Equation (9).

$$p(W \mid D) \propto p(D \mid W) \cdot p\left(W\right) \tag{9}$$

This relationship is pivotal as it refines this model's initial assumptions about the weights based on the observed data, balancing between the prior distribution and how well the model with certain weights explains the data. 'D' represents the observed dataset, making the posterior a critical component for updating beliefs about the weights after considering the data [27–31].

Predictions in a Bayesian NN are derived not from a single set of weights but by considering the entire distribution of possible weights. The predictive distribution for a new input 'x' is computed by integrating over all possible weights, each weighted by its posterior probability, Equation (10).

$$p(y \mid x, D) = \int p(y \mid x, W) p(W \mid D) dW \tag{10}$$

Since this integral is generally intractable, approximation techniques such as Monte Carlo or variational inference are employed. These methods help approximate the true predictive distribution, enabling the network to make inherently uncertain predictions and reflect the model's confidence [32–39].

2.3. Gaussian Process (GP)

The GP starts with a mean function, $m\left(x\right)$, which represents the expected value of the function at any point x, usually the mean function is set to '0'. The GP is influenced by its covariance function or kernel, $k\left(x, x^{'}\right)$. This function defines the expected covariance between any two points in the input space based on their values, which is defined as Equation (11).

$$k\left(x, x^{'}\right) = \sigma^2 exp\left(-\frac{\left(x - x^{'}\right)^2}{2l^2}\right) \tag{11}$$

Here, 'σ^2' represents the variance, controlling the variation of outputs from the mean, while l determines the length scale, influencing how input changes affect output changes. A Gaussian Process is defined by combining the mean and covariance functions, where any collection of function values follows a multivariate Gaussian distribution, Equation (12).

$$f\left(x\right) \sim GP\left(m\left(x\right), k\left(x, x^{'}\right)\right) \tag{12}$$

The predictive capability of a GP is enacted through its ability to provide expected outputs (y^*) for new inputs (X^*), using Equation (13).

$$y^* \mid X, y, X^* \sim N\left(\mu^*, \Sigma^*\right) \tag{13}$$

Here, μ^* and Σ^* are determined by the relationships established by the covariance function between known training inputs 'X', their corresponding outputs 'y', and the new input points 'X^*'. The equations for 'μ^*' and 'Σ^*' bridge the theoretical properties of the GP with

practical application, allowing the GP to forecast new values while considering both the learned patterns and inherent uncertainties,

$$\mu^* = K\left(X^*, X\right) K(X, X)^{-1} y \tag{14}$$

$$\Sigma^* = K\left(X^*, X^*\right) - K\left(X^*, X\right) K(X, X)^{-1} K\left(X, X^*\right) \tag{15}$$

In the Equation (14) and Equation (15), $K(X, X)$ is the covariance matrix computed between all pairs of training inputs, $K\left(X^*, X\right)$ is the covariance matrix between test inputs and training inputs, and $K\left(X^*, X^*\right)$ is the covariance matrix between all pairs of test inputs.

Gaussian Processes provide a robust way to predict outcomes along with a quantification of uncertainty in those predictions, which is particularly useful in fields like geo-statistics, robotics, and any domain where precise uncertainty modeling is required.

3. Methodology

3.1. Location of Study

Henan Province is a region located in the plains of China, which is among China's most agriculturally productive regions, especially in the production of wheat. The study was conducted during the period from January 2020 to December 2023. During this period, the Henan region had approximately 5 million hectares for wheat cultivation and had yielded approximately 30 million tons annually, with an average yield of around 6 tons per hectare. The yield per hectare was primarily influenced by weather conditions and farming practices. The farmers in Henan province have insured their crops based on yield- and revenue-based policies, where the government subsidizes premiums to encourage farmer participation. The LR from these insurance policies have shown fluctuations ranging from 20% to 50% over the study period, and the fluctuations are influenced mainly by climatic extremities such as droughts and floods. Further climatic pattern changes have induced farmers to adjust their agricultural practices, such as modification of sowing dates, adoption of new wheat varieties, and enhancements to the irrigation and drainage systems to better cope with changing environmental conditions.

Periodic Analysis

- **Periodic Changes**: Periodic analysis is a method used to identify and model recurring patterns in data, typically focusing on seasonal variations or trends that repeat annually or every 5–10 years.
- **Seasonality**: Monthly data can reveal seasonal patterns, such as agricultural yields, influenced by planting and harvesting cycles.

3.2. Data Sources

The data for the study include the historical loss ratio in terms of % of loss claimed against the actual insured. The data on weather factors such as rainfall, temperature fluctuations, droughts and floods are sourced from regional meteorological stations. Farming practices that include farming techniques and sowing dates are sourced from agricultural bureaus in Henan. Further economic impact reports are sourced from government reports. The following **Table 1** shows the variables, sources and data types used in this study.

Table 1. Summary of the variables, their sources, and data types.

Variable	Data Type	Data Source
Historical LR	Numeric (Percentage)	Insurance claims records
Rainfall Temperature Extreme weather events	Numeric (mm) Numeric (°C) Categorical	Regional meteorological stations
Sowing dates Wheat varieties Use of fertilizers	Date Categorical Categorical	Agricultural bureaus in Henan
Economic factors	Numeric (Currency)	Governmental reports

3.3. Data Preparation

As a first step in data preparation, the dataset was cleaned to remove inconsistencies, missing values, and outliers in the insurance claims records and weather data and crop yield figures were identified and rectified. Next, the missing data points were handled through imputation techniques that supplement the values using median values of nearby data points to maintain the integrity of the time series. To scale the data to a standard range of 0 to 1, a normalization process such as the *Min-Max* scaling technique was employed. Categorical data, such as wheat varieties and types of weather events, were transformed into numerical formats through one-hot encoding. Also, to reduce bias in economic impact, data log transformations were applied to skewed data

distributions (**Figure 3**).

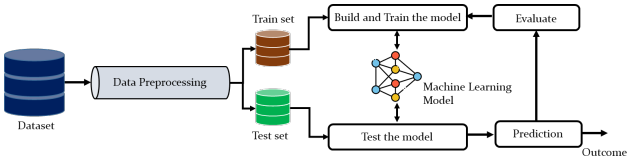

Figure 3. Prediction model.

3.4. Proposed DGP with a Bayesian LSTM for Predicting Insurance Loss Ratio

The proposed model is depicted in **Figure 3**, which integrates a DGP with a Bayesian LSTM framework. This model combines the advantage of the sequential data processing capabilities of LSTM with the probabilistic strengths of Gaussian Processes. In the data preprocessing stage, each feature is standardized to zero mean and unit variance to ensure uniformity. It is attained by the transformation of Equation (16).

$$x^{'} = \frac{x - \mu}{\sigma} \tag{16}$$

where μ and σ represent the mean and standard deviation of each feature, respectively. The Bayesian LSTM processes these normalized inputs by treating the weights within its architecture as probabilistic Gaussian distribution. Following the LSTM layer, the outputs are fed as inputs to the Deep Gaussian Process, which uses a kernel function to model the complex relationships in the data, Equation (17).

$$k\left(x, x^{'}\right) = \sigma^2 exp\left(-\frac{\|x - x^{'}\|^2}{2l^2}\right) \tag{17}$$

This Gaussian Process predicts the expected outputs and quantifies the uncertainty of these predictions using the Equation (18) and Equation (19) for the mean and covariance of the predictive distribution:

$$\mu^* = K\left(X^*, X\right) K(X, X)^{-1} y \tag{18}$$

$$\Sigma^* = K\left(X^*, X^*\right) - K\left(X^*, X\right) K(X, X)^{-1} K\left(X, X^*\right) \tag{19}$$

The model is trained to minimize the negative log-likelihood function, in which the predicted output by the model for a given input is denoted as y, with the mean μ and variance σ^2 predicted by the model, the Negative Log-Likelihood (NLL) of the Gaussian distribution is specified by Equation (20).

$$NLL = \sum_{i=1}^{N} [log\left(\sigma_i^2\right) + \frac{\left(y_i - \mu_i\right)^2}{\sigma_i^2}] \tag{20}$$

where N is the number of data points, y_i is the actual observed value, μ_i and σ_i^2 are the mean and variance predicted by the model for the ith data point. This loss function comprises two parts: the term $log\left(\sigma_i^2\right)$ ensures that the model does not become overly confident about its predictions by penalizing predictions with minimal predicted variances. $\frac{\left(y_i - \mu_i\right)^2}{\sigma_i^2}$ is essentially a weighted mean squared error that scales the squared difference between the predicted mean and the actual value by the inverse of the predicted variance, thus considering the model's uncertainty in its predictions. Once trained and validated, the model is integrated into decision-support systems, providing real-time, reliable predictions that are crucial for managing the risks associated with agricultural insurance.

Algorithm: DGP with Bayesian LSTM for Predicting Insurance LR

Inputs:

- X : Time series data of features (*e.g.,* weather conditions, agricultural practices, historical LR).
- Y : Corresponding historical LR.
- N : Number of training epochs.
- η : Learning rate for the optimizer.

Output: Predicted LR with associated confidence intervals.

Procedure:

(1) Normalize Data: $X^{'} = \frac{X - \mu_X}{\sigma_X}$, where μ_X and σ_X are the mean and standard deviation of the training data.

(2) Convert $X^{'}$ into sequences suitable for LSTM processing.

(3) Initialize Model Components:

- Bayesian LSTM Initialization: Initialize weights and biases of the LSTM layers with Gaussian distributions.
- Gaussian Process Initialization: Select a kernel function and initialize hyperparameters (length-scale l, variance σ^2).

(4) Model Training:

- For Each epoch i from 1 to N
- For Each batch (X_{batch}, Y_{batch}) in training data
- Forward Pass
- Compute LSTM outputs using the Bayesian LSTM layer.
- Feed LSTM outputs into the Gaussian Process to get predictions μ and variance σ^2.
  - Calculate Loss: $Loss = \sum_j \left[log\left(\sigma_j^2\right) + \frac{(y_j-\mu_j)^2}{\sigma_j^2}\right]$, where y_j are the actual values from Y_{batch}.
  - Backpropagation: Update the model parameters using gradient descent to minimize the loss, $Parameter\ update = Parameter - \eta \cdot \nabla (Loss)$

(5) Prediction:

- For new input data X_{new}:
  - Normalize X_{new} as done in preprocessing.
  - Process X_{new} through the trained Bayesian LSTM to get new outputs.
  - Use outputs as inputs to the Gaussian Process to predict LR with μ_{new} and σ_{new}^2.
- Output Result: Return the predicted LR, μ_{new} along with confidence intervals derived from σ_{new}^2.

(6) Model Evaluation and Adjustment:

- Evaluate model performance on a validation set.
- Adjust hyperparameters or extend training based on performance metrics such as RMSE or likelihood measures.

4. Analysis

4.1. Descriptive Analysis

The analysis results are presented in **Figure 4**, and it began with generating essential statistical summaries of the historical LR from January 2020 to December 2023. For instance, the average loss ratio over this period was initiated to be approximately 35%, with a standard deviation of 8%. The highest loss ratio observed was 50% in July 2021 due to severe flooding. A linear trend indicated a slight upward trend with an annual increase estimated at 2% per year. Seasonal decomposition revealed that LR typically peak during the late summer months, coinciding with the main harvest and the monsoon season, which brings about most of the region's rainfall and associated weather disruptions. For instance, LR were notably higher in August and September each year, consistently exceeding the annual average by 10% to 15% points. Cyclical analysis indicated a biennial pattern in the LR, with more pronounced peaks every two years. It could correlate with broader economic cycles or biennial variations in regional climate conditions. For example, 2021 and 2023 witnessed the higher peaks in LR compared to 2020 and 2022.

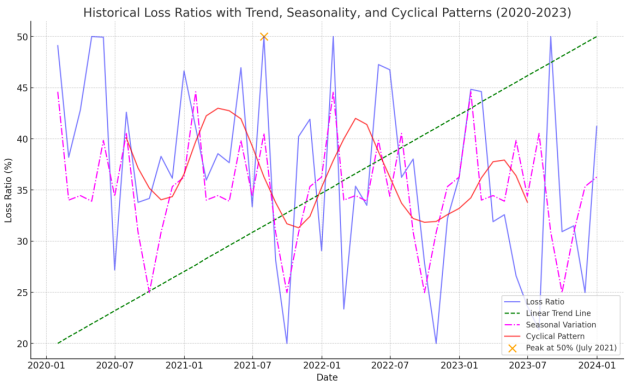

Figure 4. Statistical results summary.

The correlation analysis results are presented in **Figure 5a–c**. The analysis of LR with key factors in Henan Province revealed significant relationships: Rainfall exhibited a strong correlation coefficient of 0.48, indicating that increased precipitation correlates substantially with higher LR. Temperature showed a robust correlation at 0.82, underscoring its critical impact on LR, especially during temperature extremes. The cost of agricultural inputs had an extremely strong correlation at 0.96, reflecting that rising costs significantly influence LR, likely due to increased financial pressure on farmers. For wheat varieties, Zhengmai 366 showed almost no impact on LR (correlation of –0.013); while Xiaoyan 22 and Jinmai 47 exhibited moderate positive (0.208) and negative (–0.215) correlations, respectively, suggesting that specific varieties can affect vulnerability and resilience to conditions affecting crop yield and subsequent insurance claims.

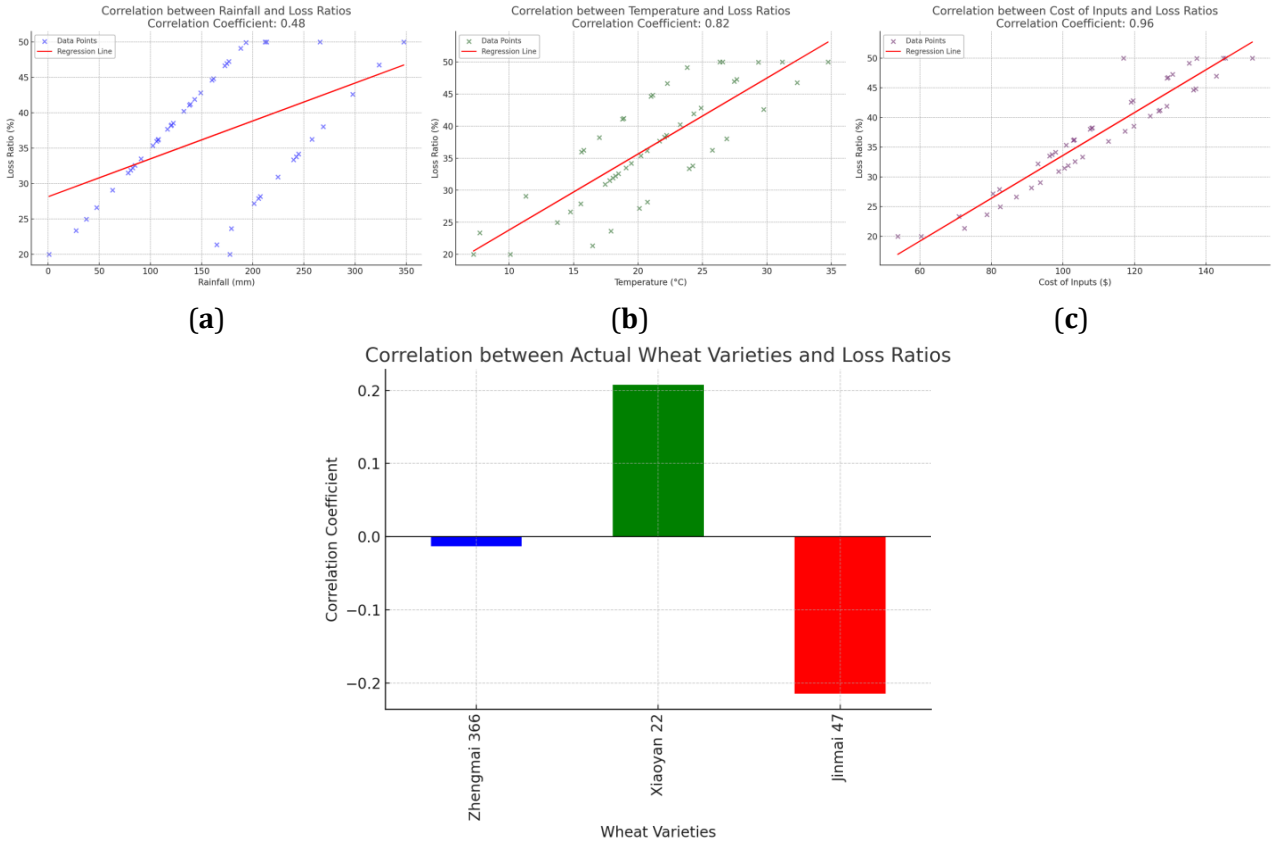

Figure 5. Correlation of (**a**) loss ratios vs. rainfall; (**b**) loss ratios vs. temperature; (**c**) loss ratios vs. cost of inputs; (**d**) Loss ratios vs. wheat varieties.

Table 2 shows Stationarity test results in which the Augmented Dickey-Fuller (ADF) test yielded a test statistic of –3.45, with a *p-value* of 0.017. Since the test statistic is supplementary negative than the critical value at 5% (–2.89), and the *p-value* is below the threshold of 0.05, we reject the null hypothesis that the series has a unit root, indicating that the series is stationary. Similarly, the Phillips-Perron (PP) test also supports the conclusion of stationarity with a test statistic of –3.60 and a *p-value* of 0.015, further reinforcing the ADF test findings by providing robustness against serial correlation and heteroscedasticity in the time series. The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, which has a null hypothesis that the series is stationary, gave a test statistic of 0.35 with a *p-value* of 0.048. As the *p-value* is marginally below 0.05, we narrowly reject the null hypothesis, suggesting that the series does not exhibit a unit root, thus confirming stationarity.

4.2. Machine Learning Analysis

The experiments were performed in a system with an Intel Core i9 processor at 3.6 GHz featuring 16 cores, which is paired with an NVIDIA RTX 3080 GPU boasting 10 GB of VRAM. The system is reinforced by 64 GB of DDR4 RAM and operates under Ubuntu 20.04 LTS. The software environment is robust, utilizing Python 3.8, TensorFlow 2.4 and GPflow 2.1 to implement and run the machine learning models. Regarding data division, 70% of the dataset is designated as the training set, and 30% is used as the test set. The proposed model was trained using the parameters listed in **Table 3**.

The proposed model is compared with Autoregressive Integrated Moving Average (ARIMA), Simple LSTM Model, Random Forest Regressor (RF), and Support Vector Regression (SVR) using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Log-Likelihood, Coefficient of Determination (R^2), Accuracy, Recall, and Area Under Curve (AUC). Each model will

Table 2. Stationarity test results for LR data.

Test	Test Statistic	p-Value	Critical Value 5%
Augmented Dickey-Fuller (ADF)	−3.45	0.017	−2.89
Kwiatkowski-Phillips-Schmidt-Shin (KPSS)	0.35	0.048	0.463
Phillips-Perron (PP)	−3.60	0.015	−2.89

be trained using the same dataset and evaluated based on the defined metrics to ensure a fair comparison (**Tables 4** and **5**).

Table 3. Training parameters.

Parameter	Value
Learning rate	0.001
Epochs	100
Batch size	32
LSTM units	100
GP kernel type	RBF
Optimizer	Adam
Early stopping criteria	Validation loss
Kernel length-scale (l)	1.0
Kernel variance (σ^2)	1.0

Tables 4 and **5** show the model's performance against different performance metrics in the training phase. The model achieved a minimum RMSE of 5.1% and a maximum of 9.3%, which is a lower error rate than other models (**Figure 6**). The mean RMSE of 7.1% was the lowest among other models like ARIMA, which has a mean RMSE of 8.4%, LSTM at 7.4% and SVR at 8.6%. For MAE, the proposed model had a lower mean value of 4.5% compared to other models' scores like ARIMA's 6.7% and SVR's 6.2%. Also, for the coefficient of determination (R^2) metric the proposed model achieved as average at 0.87, which is superior to the simpler ARIMA model's R^2 mean of 0.80 and closely approaches the theoretical maximum of 1. In the testing phase, the proposed model shows slightly increased RMSE values ranging from 7.8% to 8.6%. Yet the best model, compared to other models like the ARIMA and SVR, had values at 10.9% and 10.3 respectively, and LSTM and RF at 9% and 9.3% respectively. Also, the MAE and R^2 metrics in the testing phase show the model's efficiency with R^2 values peaking at 0.90 for the proposed model compared to 0.74 for ARIMA and 0.79 for SVR.

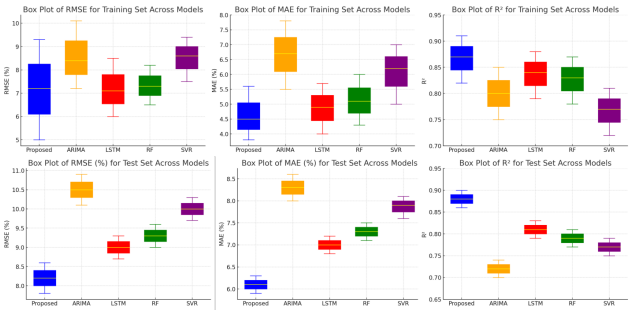

Figure 6. Boxplots for compared models for both training and test data.

The analysis of log-likelihood errors is proved in **Figure 7**, which compares different predictive models over the 100th epochs. The proposed model exhibits the most substantial improvement, starting at −250.32 and reducing to −90.45 by the 100th epoch. This suggests its superior capability in capturing complex data patterns, which is ideal for agricultural insurance loss predictions. In comparison, traditional models like ARIMA and SVR start with higher errors and show less steep improvements, indicating their relative limitations in handling the complexities of agricultural data. Simple LSTM and RF improve significantly but do not match the proposed model's efficiency.

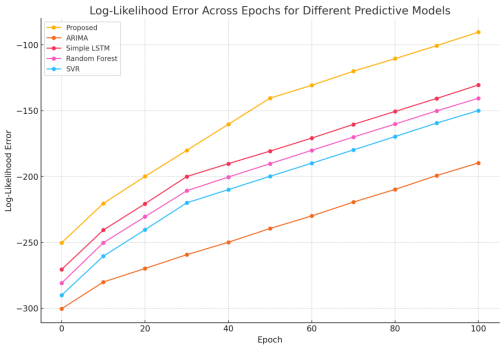

Figure 7. Error rate analysis.

The performance evaluation of various ML models in **Figure 8** for classification showcases their effectiveness across accuracy, recall, and AUC metrics on training and testing datasets.

Table 4. Results of the training test.

Model	Metric	Min Value	Mean Value	Max Value
Proposed	RMSE (%)	5.1	7.1	9.3
	MAE (%)	3.8	4.5	5.6
	R^2	0.82	0.87	0.91
ARIMA	RMSE (%)	7.2	8.4	10.1
	MAE (%)	5.5	6.7	7.8
	R^2	0.75	0.80	0.85
Simple LSTM	RMSE (%)	6.1	7.4	8.8
	MAE (%)	4.2	4.9	5.7
	R^2	0.79	0.84	0.88
Random Forest Regressor	RMSE (%)	6.5	7.3	8.2
	MAE (%)	4.3	5.1	6.0
	R^2	0.78	0.83	0.87
SVR	RMSE (%)	7.5	8.6	9.4
	MAE (%)	5.2	6.2	7.0
	R^2	0.72	0.77	0.81

Table 5. Results for test data.

Model	Metric	Min Value	Mean Value	Max Value
Proposed	RMSE (%)	7.8	8.2	8.6
	MAE (%)	5.9	6.1	6.3
	R^2	0.86	0.88	0.90
ARIMA	RMSE (%)	10.1	10.5	10.9
	MAE (%)	8.2	8.3	8.6
	R^2	0.70	0.72	0.74
Simple LSTM	RMSE (%)	8.7	9.0	9.3
	MAE (%)	6.8	7.0	7.2
	R^2	0.79	0.81	0.83
Random Forest Regressor	RMSE (%)	9.1	9.3	9.6
	MAE (%)	7.1	7.3	7.5
	R^2	0.77	0.79	0.81
SVR	RMSE (%)	9.7	10.0	10.3
	MAE (%)	7.6	7.9	8.1
	R^2	0.75	0.77	0.79

The proposed model demonstrates exemplary performance on the training set, achieving an accuracy of 0.99, a recall of 0.97, and an AUC of 0.97, indicating its near-perfect ability to classify correctly and manage class imbalances. On the test set, it maintains robust performance with an accuracy of 0.97, recall of 0.84, and AUC of 0.92, showing a slight decrease but still outperforming other models. The RF model also shows strong results, with training metrics of 0.98 accuracy, 0.92 recall, and 0.96 AUC, and testing metrics close behind at 0.96 accuracy, 0.83 recall, and 0.91 AUC. This reflects its effective generalization with a minor performance drop on new data. LSTM, while less effective than the proposed and RF models, achieves a training accuracy of 0.93, recall of 0.69, and AUC of 0.84, which decreases in the testing phase to 0.87 accuracy, 0.50 recall and 0.73 AUC. These results suggest its relative weakness in handling this classification task, especially in maintaining recall of unseen data. SVR and ARIMA models exhibit similar patterns, scoring training and testing accuracies of 0.90. However, both struggle with recall and AUC metrics, registering 0.50 on recall and 0.75 on AUC across both datasets, indicating limitations in their classification effectiveness and sensitivity compared to other models.

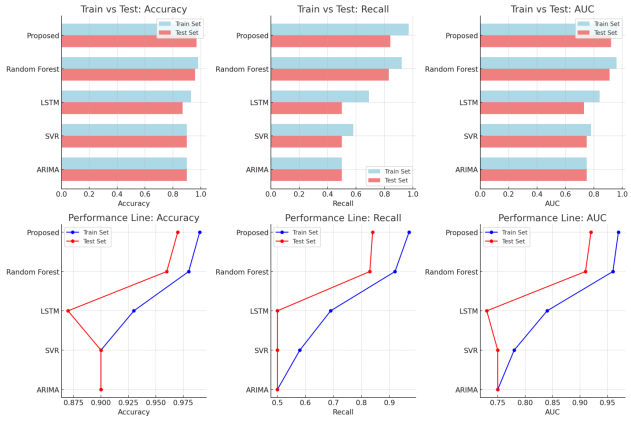

Figure 8. Performance of classification ML models evaluated by accuracy, recall rate and AUC.

The Execution Time (ET) results in **Figure 9** compare the efficiency of various ML models during training and prediction.

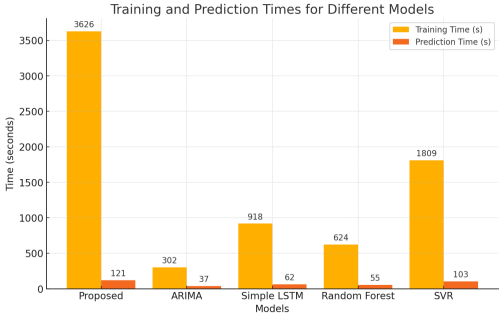

Figure 9. Execution Time analysis.

The proposed model, featuring complex computations, has the longest training time at 3626 *sec.* However, it maintains a moderate prediction time of 121 *sec.*, demonstrating its capacity to handle complexity efficiently during real-time applications. ARIMA stands out for its efficiency, with the shortest training and prediction times at 302 and 37 *sec.* respectively, making it exceptionally quick and suitable for rapid output needs. The Simple LSTM and Random Forest models exhibit moderate training times of 918 and 624 *sec.*, with relatively quick prediction times of 62 and 55 *sec.* respectively, aligning them well with scenarios that balance training depth with prompt predictions. The SVR requires 1809 *sec.* for training and 103 *sec.* for predictions that show moderate computational demands compared to the other models.

The Computational Complexity for each model is listed in **Table 6**, and the proposed model has $O\left(n^3\right)$ complexity, which is the most resource-intensive; whereas the ARIMA, LSTM, and Random Forest models display moderate complexities of $O\left(nk^2\right)$, $O(t \cdot p)$ and $O(t \cdot m \cdot log(n))$ respectively. The SVR, similar to the proposed model, shows high computational demands with a complexity that ranges from $O\left(n^2 \cdot p\right)$ to $O\left(n^3\right)$.

Table 6. Computational Complexity analysis.

Model	**Time Complexity**
Proposed	High $\left(O\left(n^3\right)\right)$
ARIMA	Moderate $\left(O\left(nk^2\right)\right)$
Simple LSTM	Moderate $(O(t \cdot p))$
Random Forest	Moderate $(O(t \cdot m \cdot log(n)))$
SVR	High $\left(O\left(n^2 \cdot p\right)\right.$ to $\left.O\left(n^3\right)\right)$

5. Conclusions and Future Work

Crop insurance has been a life savior for many farmers by mitigating the financial risks arising from different climatic and environmental uncertainties. The study was conducted in Henan Province, China, and data related to historical insurance claims, weather data, and government policies were collected. The study proposed a Deep Gaussian Process with Bayesian Long Short-Term Memory networks to predict insurance LR accurately. This model combines the advantage of the sequential data processing capabilities of LSTM with the probabilistic strengths of Gaussian Processes. The model was trained and tested using the dataset sourced from a period from January 2020 to December 2023. The results of the experiment have shown that the proposed model has better prediction accuracy than the other traditional models like LSTM, ARIMA, RF and SVR.

The research aims to evaluate the effectiveness of Long Short-Term Memory (LSTM) models in predicting agricultural outcomes or insurance mechanisms in the context of climate change. It compares LSTMs' predictive capabilities with other statistical and machine learning models, and explores their potential contributions to the field.

To strengthen the research, it is crucial to clearly communicate its significance.

- Climate change is significantly impacting agricultural production and insurance mechanisms, necessitating accurate prediction models due to its

impact on crop yields, pest prevalence, and the frequency of extreme weather events.

- The text discusses the comparison of LSTM against traditional models like ARIMA or linear regression, as well as machine learning approaches like Random Forests and Gradient Boosting Machines, and their strengths and limitations in time-series forecasting for agricultural data.
- LSTM models are ideal for capturing temporal patterns in climate and agricultural data due to their effectiveness in long-term dependencies in data sequences.
- The research should highlight its contribution to previous studies and identify any gaps in the literature, such as examining the effectiveness of LSTM models in prediction accuracy or robustness compared to earlier models.
- The study highlights the potential benefits of improved prediction accuracy for agricultural producers and insurance companies, such as accurate risk assessments, optimized resource allocation, and improved decision-making.
- The text details the data used, the methodology for evaluating model performance, and metrics like MAE and RMSE, which are relevant to the research.

Future research will focus on extending this approach by incorporating real-time data and exploring its efficacy in different geographic regions, thereby offering broader implications for stabilizing agricultural insurance markets under dynamic environmental conditions.

Future research should explore hybrid models combining LSTMs with other techniques, or apply findings to diverse regions or crop types.

Author Contributions

Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization: Y.W., M.A.B.A., J.Y.T.H.

Funding

This work received no funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflict of Interest

All authors disclosed no conflict of interest.

References

[1] DeBoe, G., 2020. Economic and environmental sustainability performance of environmental policies in agriculture. OECD Food, Agriculture and Fisheries Papers. Paper no. 140. OECD Publishing, Paris.

[2] King, M., Singh, A.P., 2020. Understanding farmers' valuation of agricultural insurance: Evidence from Vietnam. Food Policy. 94, 101861.

[3] Zhong, L., Nie, J., Yue, X., et al., 2023. Optimal design of agricultural insurance subsidies under the risk of extreme weather. International Journal of Production Economics. 263, 108920.

[4] Möhring, N., Dalhaus, T., Enjolras, G., et al., 2020. Crop insurance and pesticide use in European agriculture. Agricultural Systems. 184, 102902.

[5] Chowdhury, S., Mayilvahanan, P., Govindaraj, R., 2022. Optimal feature extraction and classification-oriented medical insurance prediction model: Machine learning integrated with the internet of things. International Journal of Computers and Applications. 44(3), 278–290.

[6] Dhieb, N., Ghazzai, H., Besbes, H., et al., 2020. A secure AI-driven architecture for automated insurance systems: Fraud detection and risk measurement. IEEE Access. 8, 58546–58558.

[7] Richman, R., 2021. AI in actuarial science—A review of recent advances—Part 2. Annals of Actuarial Science. 15(2), 230–258.

[8] Manteigas, C., António, N., 2024. Understanding and predicting lapses in mortgage life insurance using a machine learning approach. Expert Systems with Applications. 255(Part C), 124753.

[9] Quan, Z., Hu, C., Dong, P., et al., 2024. Improving business insurance loss models by leveraging InsurTech innovation. arXiv preprint. arXiv:2401.16723.

[10] Cravero, A., Pardo, S., Galeas, P., et al., 2022. Data type and data sources for agricultural big data and machine learning. Sustainability. 14(23), 16131.

[11] Chai, C., Zhang, B., Li, Y., et al., 2023. A new multi-dimensional framework considering environmental impacts to assess green development level of cultivated land during 1990 to 2018 in China. Environmental Impact Assessment Review. 98, 106927.

[12] Fernando, N., Kumarage, A., Thiyaganathan, V., et al. (editors), 2022. Automated vehicle insurance claims processing using computer vision, natural language processing. 2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer); 30 November–1 December 2022; Colombo, Sri Lanka. pp. 124–129. DOI: https://doi.org/10.1109/ICTer58063.2022.10024089

[13] Ramalingam, H., Venkatesan V.P., 2019. Conceptual analysis of Internet of Things use cases in Banking domain. TENCON 2019—2019 IEEE Region 10 Conference (TENCON); 12 December 2019; Kochi, India. pp. 2034–2039. DOI: https://doi.org/10.1109/TENCON.2019.8929473

[14] Martin, J.M.R., 2021. Designing and verifying microservices using CSP. 2021 IEEE Concurrent Processes Architectures and Embedded Systems Virtual Conference (COPA); 23 September 2021; San Diego, CA. pp. 1–4. DOI: https://doi.org/10.1109/COPA51043.2021.9541471

[15] Cardoso, J., 2006. Benchmarking a semantic Web service srchitecture for fault-tolerant B2B integration. 26th IEEE International Conference on Distributed Computing Systems Workshops (ICDCSW'06); 24 July 2006; Lisboa, Portugal. p. 18. DOI: https://doi.org/10.1109/ICDCSW.2006.27

[16] Romania, J., Ross, W., Butcher, S., 2017. Army and Navy management of Automatic Test Systems for weapon system support: Comparing US Army and US navy ATS management practices. IEEE AUTOTESTCON; 26 October 2017; Schaumburg, IL. pp. 1–9. DOI: https://doi.org/10.1109/AUTEST.2017.8080459

[17] Benrachou, D.E., Glaser, S., Elhenawy, M., et al., 2024. Improving efficiency and generalisability of motion predictions with deep multi-agent learning and multi-head attention. IEEE Transactions on Intelligent Transportation Systems. 25(6), 5356–5373. DOI: https://doi.org/10.1109/TITS.2023.3339640

[18] Agaram, M., 2018. Intelligent discovery features for EDM and MDM systems. 2018 IEEE 22nd International Enterprise Distributed Object Computing Workshop (EDOCW); 16–19 October 2018; Stockholm, Sweden. pp. 135–144. DOI: https://doi.org/10.1109/EDOCW.2018.00028

[19] Rahmani, M.K.I., Ghanimi, H.M., Jilani, S.F., et al., 2023. Early Pathogen Prediction in Crops Using Nano Biosensors and Neural Network-Based Feature Extraction and Classification. Big Data Research. 34, 100412. DOI: https://doi.org/10.1016/j.bdr.2023.100412

[20] Krishnamoorthy, P., Satheesh, N., Sudha, D., et al., 2023. Effective Scheduling of Multi-Load Automated Guided Vehicle in Spinning Mill: A Case Study. IEEE Access. 11, 9389–9402. DOI: https://doi.org/10.1109/ACCESS.2023.3236843

[21] Sabry, E.S., Elagooz, S., El-Samie, F.E.A., et al., 2022. Sketch-Based Retrieval Approach Using Artificial Intelligence Algorithms for Deep Vision Feature Extraction. Axioms. 11(12), 663. DOI: https://doi.org/10.3390/axioms11120663

[22] Roque-Claros, R.E., Flores-Llanos, D.P., Maquera-Humpiri, A.R., et al., 2024. UAV Path Planning Model Leveraging Machine Learning and Swarm Intelligence for Smart Agriculture. Scalable Computing: Practice and Experience. 25(5), 3752–3765. DOI https://doi.org/10.12694/scpe.v25i5.3131

[23] Ghanimi, H.M., Suguna, R., Jeyaraj, J.P.G., et al., 2024. Smart Fertilizing Using IOT Multi-Sensor and Variable Rate Sprayer Integrated UAV. Scalable Computing: Practice and Experience. 25(5), 3766–3777. DOI: https://doi.org/10.12694/scpe.v25i5.3132

[24] Nowfal, S.H., Sadu, V.B., Sengan, S., et al., 2024. Genetic Algorithms for Optimized Selection of Biodegradable Polymers in Sustainable Manufacturing Processes. Journal of Machine and Computing. 4(3), 563–574. DOI: https://doi.org/10.53759/7669/jmc202404054

[25] Jeevika Tharini, V., Ravi Kumar, B., Sahaya Suganya Princes, P., et al., 2024. Business Decision-Making Using Hybrid LSTM for Enhanced Operational Efficiency. In: Vimal, V., Perikos, I., Mukherjee, A., et al. (eds.) Multi-Strategy Learning Environment. ICMSLE 2024. Algorithms for Intelligent Systems. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-97-1488-9_12

[26] Jermanshiyamala, A., Kumar, N.S., Belhe, S., et al., 2024. ACO-Optimized DRL Model for Energy-Efficient Resource Allocation in High-Performance Computing. In: Vimal, V., Perikos, I., Mukherjee, A., et al. (eds.) Multi-Strategy Learning Environment. ICMSLE 2024. Algorithms for Intelligent Systems. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-97-1488-9_11

[27] Nowfal, S.H., Rao, G.R.K., Velmurugan, V., et

al., 2024. Advancing viscoelastic material modeling: Tackling time-dependent behavior with fractional calculus. Journal of Interdisciplinary Mathematics. 27(2), 307–316. DOI: https://doi.org/10.47974/JIM-1827

[28] Vidya Sagar, P., Rajyalaxmi, M., Subbalakshmi, A.V.V.S., et al., 2024. Utilizing stochastic differential equations and random forest for precision forecasting in stock market dynamics, Journal of Interdisciplinary Mathematics. 27(2), 285–298. DOI: https://doi.org/10.47974/JIM-1822

[29] Lazar, A.J.P., Soundararaj, S., Sonthi, V.K., et al., 2023. Gaussian Differential Privacy Integrated Machine Learning Model for Industrial Internet of Things. SN Computer Science. 4, 454. DOI: https://doi.org/10.1007/s42979-023-01820-2

[30] Mehbodniya, A., Webber, J.L., Mani, D., et al., 2022. Classification of Cervical Cells Using Deep Learning Feature Extraction, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems. Springer, Singapore. Volume 565, pp. 27–41. DOI: https://doi.org/10.1007/978-981-19-7455-7_3

[31] Karn, A.L., Webber, J.L., Mehbodniya, A., et al., 2022. Evaluation and Language Training of Multinational Enterprises Employees by Deep Learning in Cloud Manufacturing Resources, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems. Springer, Singapore. Volume 565, pp. 369–380. DOI: https://doi.org/10.1007/978-981-19-7455-7_28

[32] Karn, A.L., Mehbodniya, A., Webber, J.L., et al., 2022. Design of Concurrent Engineering Systems for Global Product Development Using Artificial Intelligence, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems. Springer, Singapore. Volume 565, pp. 425–434. DOI: https://doi.org/10.1007/978-981-19-7455-7_32

[33] James, G.M.B., Mehbodniya, A., Maria, A.B., et al., 2022. Deep Convolutional Neural Networks-Based Market Strategy for Early-Stage Product Development, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems. Springer, Singapore. Volume 565, pp. 597–606. DOI: https://doi.org/10.1007/978-981-19-7455-7_46

[34] Bhavana Raj, K., Webber, J.L., Marimuthu, D., et al., 2022. Equipment Planning for an Automated Production Line Using a Cloud System, Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems. Springer, Singapore. Volume 565, pp. 707–717. DOI: https://doi.org/10.1007/978-981-19-7455-7_57

[35] Mathew, T.E., Sabu, A., Sengan, S., et al., 2023. Microclimate monitoring system for irrigation water optimization using IoT. Measurement: Sensors. 27, 100727. DOI: https://doi.org/10.1016/j.measen.2023.100727

[36] Gupta, N.V.R., Rajeshkumar, G., Selvi, S.A.M., et al., 2022. Li-Fi Enables Reliable Communication of VLT for Secured Data Exchange, Intelligent Systems and Sustainable Computing. Smart Innovation, Systems and Technologies. Springer, Singapore. Volume 289. DOI: https://doi.org/10.1007/978-981-19-0011-2_20

[37] Mantena, S.V., Jayasundar, S., Sharma, D.K., et al., 2022. Design of Dual-Stack, Tunneling, and Translation Approaches for Blockchain-IPv6, Intelligent Systems and Sustainable Computing. Smart Innovation, Systems and Technologies. 289. DOI: https://doi.org/10.1007/978-981-19-0011-2_21

[38] Sengan, S., Khalaf, O.I., Ettiyagounder, P., et al., 2022. Novel Approximation Booths Multipliers for Error Recovery of Data-Driven Using Machine Learning, Communications in Computer and Information Science. International Conference on Emerging Technology Trends in Internet of Things and Computing, TIOTC 2021: Emerging Technology Trends in Internet of Things and Computing. Springer, Cham. pp. 299–309. DOI: https://doi.org/10.1007/978-3-030-97255-4_22

[39] Dadheech, P., Sheeba, R., Vidya, R., et al., 2020. Implementation of Internet of Things-Based Sentiment Analysis for Farming System. Journal of Computational and Theoretical Nanoscience. 17(12), 5339–5345. DOI: https://doi.org/10.1166/jctn.2020.9426