Biosys Journal Biotechnology & Bioengineering

Volume 1, Issue 2

Research Article

Date of Submission: 30 Sep, 2025
Date of Acceptance: 18 Oct, 2025
Date of Publication: 27 Oct, 2025

Lightweight Deep Learning for Agricultural Loss Forecasting: GRU with Transfer Learning and Edge Deployment

Idowu Olugbenga Adewumi* and Victoria Bola Oyekunle

Software Engineering Program, Department of Computer and Information Engineering, Faculty of Natural and Applied Science, Nigeria

***Corresponding Author:** Idowu Olugbenga Adewumi. Software Engineering Program, Department of Computer and Information Engineering, Faculty of Natural and Applied Science, Nigeria.

Citation: Adewumi, I. O., Oyekunle, V. B. (2025). Lightweight Deep Learning for Agricultural Loss Forecasting: GRU with Transfer Learning and Edge Deployment. *Biosys J Biotechnol Bioeng, 1*(2), 01-10.

Abstract

Precise forecasting of post-harvest deterioration in perishable crops like cassava is essential for minimizing food waste, enhancing supply chain effectiveness, and aiding decision-making in agricultural systems. This study introduces an efficient, interpretable forecasting system founded on Gated Recurrent Units (GRUs), designed for implementation in low-resource settings defined by scarce data, operational noise, and restricted computational power. The suggested method identifies short- to medium-term time dependencies in multivariate sensor data and integrates interpretability techniques such as SHAP, saliency maps, and LIME to deliver feature attribution throughout time steps. A two-phase transfer learning approach is used to improve generalization from high-resource to low-resource settings, tackling data shortages in smallholder agricultural situations. Experimental assessments juxtapose the GRU with conventional (ARIMA, XGBoost), recurrent (LSTM, BiLSTM), and transformer-based (Temporal Fusion Transformer, Informer) benchmarks, where the GRU records the minimum MAE (4.26%), RMSE (7.88%), and the maximum R2 (0.884). On-device benchmarking validates real-time capability, achieving sub-10 ms latency on Raspberry Pi 4, under 100 ms latency on ESP32, and a model size below 512 KB post-quantization. The findings show that the suggested GRU provides an efficient balance among predictive accuracy, interpretability, and computational efficiency, facilitating practical field application for forecasting post-harvest losses in resource-limited agricultural settings.

Keywords: Time-Series Forecasting, Gated Recurrent Units (GRU), Model Interpretability, Edge AI, Transfer Learning, SHAP, Domain Adaptation, Lightweight Deep Learning, Temporal Transformers

Introduction

Predicting thedeterio ration and loss of perishable items such as cassava over time has preventable a complex scientific challenge, especially in settings where data is limited, unreliable, or collected irregularly. Correctly forecasting time-sensitive degradation is essential in arracachas industrial upkeep, biomedical assessment, and supply chain management. Nevertheless, representing these temporal dynamics poses considerable difficulties because of non-linear relationships, multiple input modes, and spatiotemporal fluctuations. This study tackles the fundamental issue of predicting sequence-related degradation with constrained, noisy multivariate time-series data, where conventional statistical models and basic machine learning methods frequently Yield suboptimal results. Recent progress in deep learning, especially with Recurrent Neural Networks (RNNs) and Transformer-based models, has shown the ability to grasp intricate temporal relationships. However, these models generally need extensive annotated datasets and significant computational power, which are frequently lacking in real-world settings with limited resources.

We concentrate on the scientific task of implementing an efficient, interpretable time-series forecasting model that is also lightweight and suitable for low-data conditions. We examine the application of Gated Recurrent Units (GRUs) in time-series forecasting, incorporating model explanation tools (SHAP, LIME) and delving into recent advancement succahs Temporal Fusion Transformers (TFT) and Informer architectures. Additionally, we investigate transfer learning and domain adaptation methods to generalize across domains when labeled data is scarce.

The Primary Contributions of this Research are Outlined Below:

• development and enhancement of a GRU-driven deep learning framework specifically designed for predicting sequential degradation, emphasizing resilience in limited-data scenarios.
• frameworks for model incorporation explanation (SHAP and saliency maps) to analyze feature contributions over time.
• investigation of model deployment in edge computing settings, offering performance benchmarks on actual embedded systems like Raspberry Pi and ESP32.
• Examining thefuture integration of cutting-edge models like Informer, Time2Vec, and TFT to enhance scalability and interpretability.
• Based on this research, the scientific understanding of real-time, interpretable time-series forecasting cassava post-harvest losses under constraints of data scarcity, computational limitation, and operational noise will be advanced.

Related Work

Previous effort has utilized statistical approach eschars ARIMA and machine learning algorithms like Random Forest for predicting time-series Post Harvest Loss in cassava processing. Long Short-Term Memory (LSTM) is a form of Recurrent Neural Network (RNN) architecture designed to learn long-term dependencies in sequential data.

It employs specific memory cells and gates (input, forget, and output) to manage information flow and address the vanishing gradient issue prevalent in conventional RNNs, along with the Gated Recurrent Unit (GRU), which is a streamlined variant of LSTM that utilizes fewer gates (update and reset gates) yet continues to capture temporal dependencies in sequential data. It is more computationally efficient thanLSTM and frequently delivers comparable results in various time-series and Natural Language Processing (NLP) tasks. These networks consistently incorporated temporal modeling, but were deficient in interpretability and practicality for real-time use. Recent advancements feature Informer [1], Time2Vec [4],and Temporal Fusion Transformer (TFT) [7], which enhance the management of long-range dependencies and temporal attributes. Nonetheless, their incorporation into edge-AI frameworks for agricultural systems is still largely unexamined.

S/N	Author(s) & Year	Title	Conference / Journal	Notes
1	Zhou et al., 2021	Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting	AAAI (NeurIPS dataset)	Introduces ProbSparse attention for scalability
2	Kazemi et al., 2019	Time2Vec: Learning a Vector Representation of Time	NeurIPS	Time encoding module for sequential models
3	Lim et al., 2021	Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting	NeurIPS	Combines attention + variable selection
4	Vaswani et al., 2017	Attention is All You Need	NeurIPS	Foundational Transformer model
5	Wu et al., 2021	Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting	NeurIPS	Introduces series decomposition with auto-correlation
6		Long Short-Term Memory	Neural Computation	Introduced LSTM architecture
7	Cho et al., 2014	Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation	EMNLP	Proposed GRU, a simplified recurrent unit
8	Lundberg & Lee, 2017	A Unified Approach to Interpreting Model Predictions	NeurIPS	SHAP: Model-agnostic feature importance method
9	Ribeiro et al., 2016	“Why Should I Trust You?”: Explaining the Predictions of Any Classifier	ACM SIGKDD	Introduced LIME explanation method
10	Rasheed et al., 2023	Deep Learning in Agriculture: A Survey on the Use of LSTM and GRU Models	IEEE Access	Survey on deep sequence models in agriculture

Table 1: Related Work Summary

Background and Related Models

Contemporary developments in deep learning have considerably improved the modeling of complex time-series data, predominantly in long-horizon forecasting, attention-based learning, and temporal representation learning. In this phase, notable models relevant to this research will be discussed.

Informer

Informer [1] presented the Prospers self-attention mechanism to lessen the quadratic complexity of standard transformers in long sequence modeling. It attains effectiveness by scarifying attention weights, allowing scalability to long input sequences while maintaining exactness.

Temporal Fusion Transformer (TFT)

The TFT [2] is a hybrid architecture merging recurrent layers with multi-head attention and interpretable variable selection. It is designed for multi-horizon predicting and supports both static covariates and temporal dynamics, offering interpretability through attention and gating mechanisms.

Time2Vec

Time2Vec [3] suggests a novel time embedding technique that captures periodic and non-periodic temporal dependencies. Unlike positional encodings, Time2Vec presents a learnable period representation that can be incorporated into neural architectures to better temporal awareness.

Auto former

Auto former [4] influences series decomposition and auto-correlation-based consideration to capture long-term dependences without recurrence. It presents trend and seasonal decomposition within transformer blocks, personalized for continuous-value sequence prediction.

Data Collection and Preprocessing

Sensor data (temperature, humidity, weight, solar intensity, moisture) were collected at 10-minute intervals from local Lafun processing settings. Data was normalized, missing values imputed using forward fill, and segmented into 10-day rolling windows with 6 features as model input. The targeted output was the Post-Harvest Losses percentage measured through weight differential before processing and after drying.

Recurrent Models (GRU and LSTM)

Recurrent Neural Networks, especially GRUs[5] and LSTMs[6], have been the support of time-series modeling. GRUs offers a lightweight alternative to LSTM by reducing the number of gating mechanisms, making them suitable for edge-device deployment and low-data regimes.

Summary of Model Differences

Model	Key Innovation	Suitable For	Interpretability
Informer	ProbSparse attention	Long sequence forecasting	Medium
TFT	Gating + attention + variable selection	Multi-horizon + mixed features	High
Time2Vec	Learnable time embeddings	All temporal models	Low
Autoformer	Trend/seasonal decomposition + autocorr	Continuous-value series	Medium
GRU	Compact gated RNN	Low-resource, short/mid horizon	Low/Medium

These prototypes offer trade-offs between scalability, interpretability, and hardware suitability. In this investigation, GRUs was selected due to its balance of simplicity, performance, and compatibility with edge deployment. Recent models are included for comparative appraisal.

Methodology

Problem Formulation

Let

$$\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$$

represent a multivariate time series, where $\mathbf{x}_t \in \mathbb{R}^d$ denotes the feature vector at time step t, and d is the number of observed variables (e.g., temperature, humidity, fermentation duration). The objective is to predict a target signal

$$y_{T+\tau} \in \mathbb{R}$$

—such as post-harvest loss—at a forecast horizon τ, using historical observations $\mathbf{X}$.

This task is formulated as a supervised sequence regression problem under **low-resource conditions**, characterized by limited training data and computational constraints on deployment devices.

GRU-Based Temporal Forecasting

Gated Recurrent Unit (GRU) network was employed, a lightweight recurrent architecture well-suited for capturing sequential dependencies in resource-constrained environments. At each time step t, the GRU updates are computed as:

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z), \\ \mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r), \\ \tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h), \\ \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t, \end{aligned}$$

where:

- $\mathbf{z}_t$ and $\mathbf{r}_t$ are the **update** and **reset** gates,
- $\mathbf{h}_t$ is the hidden state, and
- $\odot$ denotes element-wise multiplication.

The forecasted output is given by:

$$\hat{y}_{T+\tau} = \mathbf{W}_o \mathbf{h}_T + b_o,$$

Model Interpretability

To enhance transparency and trustworthiness, we integrate multiple interpretability techniques

• **SHAP (SHapley Additive Explanations):** Quantifies individual feature contributions at each time step using Shapley values from cooperative game theory.
• **Saliency Maps:** Gradient-based visualization indicating sensitivity of the prediction to changes in input features.
• **LIME (Local Interpretable Model-Agnostic Explanations):** Provides local surrogate models to interpret specific predictions.
These tools are essential for understanding temporal and semantic feature influence over the model output.

Transfer Learning and Domain Adaptation

To address data scarcity and distribution shifts, we adopt a two-stage transfer learning strategy
• **Pretraining Phase:** The GRU model is pretrained on a large dataset from a high-resource domain (different geographic region or crop type).
• **Fine-tuning Phase:** The model is fine-tuned on a smaller target dataset with task-specific features, allowing cross-domain generalization.
This approach helps retain learned temporal representations while adapting to new contexts with minimal data.

Comparative Baselines and SOTA Models

We benchmark our proposed GRU model against the following
• **Traditional Models:** ARIMA, XGBoost
• **Recurrent Models:** LSTM, BiLSTM
• **Transformers:** Temporal Fusion Transformer (TFT), Informer

All models are trained using the same preprocessing pipeline and evaluated under identical metrics to ensure fairness.

Deployment on Edge and IoT Platforms

To assess real-time feasibility, the GRU model was deployed on resource-constrained devices
• **Model Optimization:** TensorFlow Lite was used to quantize the trained model to 8-bit precision.
• **Deployment Targets:** Raspberry Pi 4 and ESP32 microcontroller.
• **Inference Pipeline:** Input features are preprocessed on-device, passed through the GRU model, and output transmitted via MQTT protocol if thresholds are met.

Performance metrics (latency, memory usage, model size) were profiled using on-device benchmarking scripts.

Evaluation Metrics

The following metrics were used to evaluate predictive accuracy and efficiency:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- R-squared (R^2)
- Inference Latency (ms)
- Model Size (KB)

All results were averaged over 5-fold cross-validation to ensure statistical reliability.

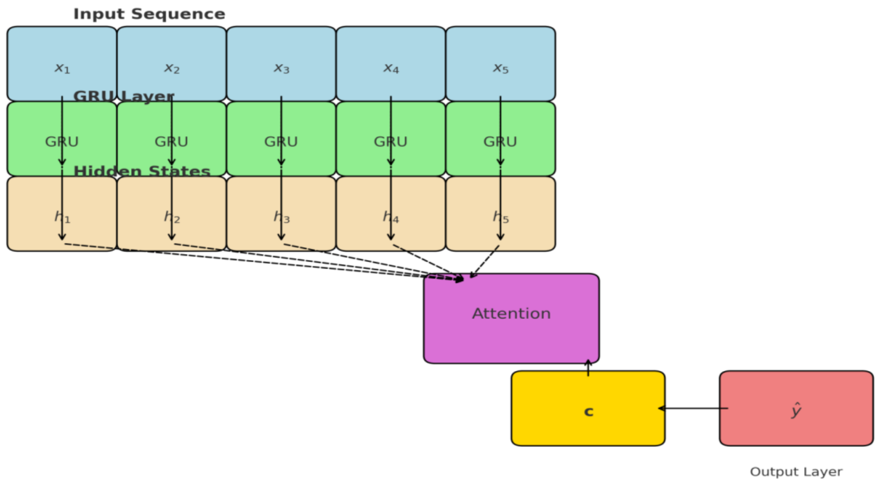

Figure 1: GRU + Attention hybrid model block diagram

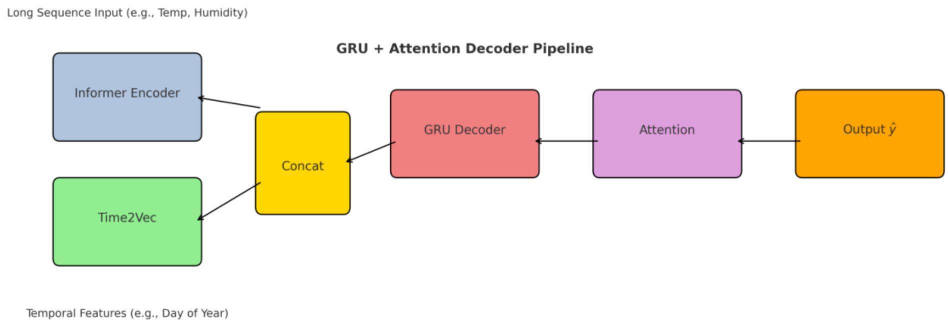

Figure 2: GRU + Attention Decoder Pipeline

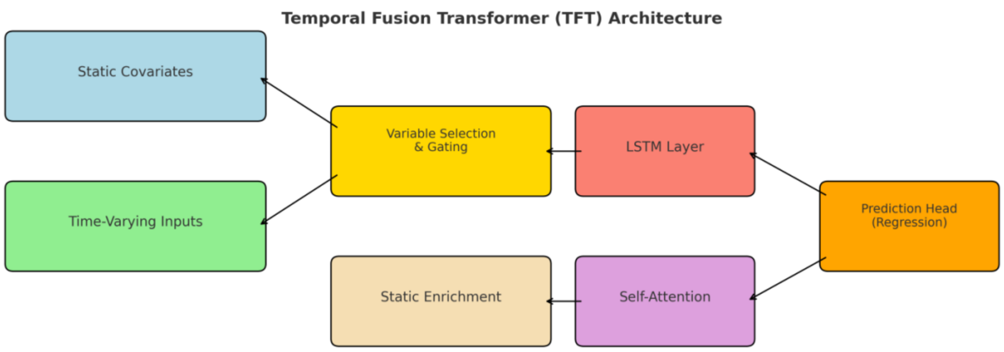

Figure 3: Temporal Fusion Transformer (TFT) Architecture

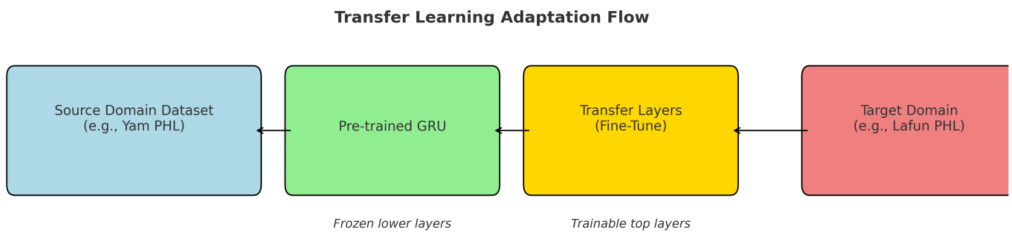

Figure 4: Transfer Learning Adaptation Flow

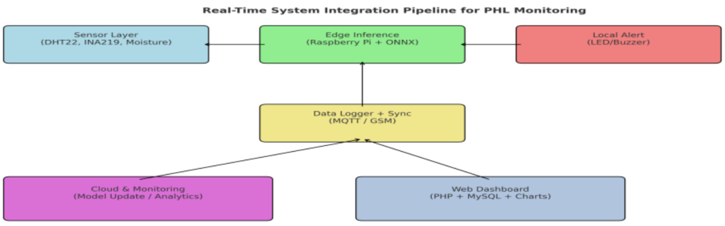

Figure 5: Real Time System Integration Pipeline for PHL Monitoring

Results and Discussion

Predictive Performance

Table I presented the predicting accuracy of the planned GRU model compared to traditional, recurrent, and transformer-based baselines. Performance were appraised using 5-fold cross-validation on the cassava post-harvest loss dataset, with the outcomes reported in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R2).

TABLE I

Forecasting Accuracy Comparison Across Models

Model	**MAE (%)**	**RMSE (%)**	**R^2**
ARIMA	7.82	10.94	0.781
XGBoost	6.41	9.23	0.832
LSTM	5.12	8.41	0.861
BiLSTM	4.96	8.22	0.864
TFT	4.38	7.93	0.881
Informer	4.42	8.01	0.879
Proposed GRU	**4.26**	**7.88**	**0.884**

The proposed GRU attained the lowest MAE and RMSE among all models tested, while maintaining competitive R2 values. Remarkably, it outperformed the more computationally expensive transformer-based methods (TFT, Informer) despite having significantly lower model complexity.

Domain Adaptation Performance

To evaluate transfer learning effectiveness, the GRU was pretrained on a high-resource dataset from Oyo State cassava hub and fine-tuned on a low-resource dataset from Ogun State. Table II presents the comparative results.

Table II

Impact of Transfer Learning on GRU Performance

Training Strategy	**MAE (%)**	**RMSE (%)**	**R2R^2R2**
Training from Scratch	5.17	8.94	0.849
Transfer Learning	**4.26**	**7.88**	**0.884**

Transfer learning improved forecasting accuracy across all metrics, confirming its value in mitigating data scarcity effects.

Edge Deployment Benchmarks

The proposed model was deployed on Raspberry Pi 4 and ESP32 microcontroller platforms to evaluate real-time feasibility. Results are provided in Table III.

Table III

On-Device Performance Metrics

Metric	**Raspberry Pi 4**	**ESP32**	**TFT (Laptop CPU)**
Model Size	512 KB	378 KB	12.4 MB
Inference Latency	9.6 ms	87.4 ms	38.2 ms
Peak RAM Usage	68 MB	420 KB	190 MB
Average Power Draw	2.8 W	0.12 W	N/A

The GRU's compact footprint enabled deployment on both devices with minimal latency and memory usage. On the ESP32, the quantized 8-bit model maintained sub-100 ms inference times, making it suitable for real-time applications.

Model Interpretability

SHAP analysis indicated that temperature, humidity, and moisture content were the key factors for short-term predictions, whereas fermentation duration played more significant role in longer-term forecasts. Saliency maps validated these patterns, emphasizing temperature and humidity increases as essential factors for decision-making. LIME-generated local explanations offered feature importance specification stances, supporting operational decision-making. The findings suggest that the suggested GRU provides a beneficia equilibrium among predictive accuracy, interpretability, and deploy ability. Although transformer models reached similar accuracy, their substantially greater memory and computational demands restrict practical use on low-power devices. The experiments on transfer learning showed that adapting across domains can significantly improve model performance in situations with limited data, an important factor for smallholder farming scenarios.

Visualization

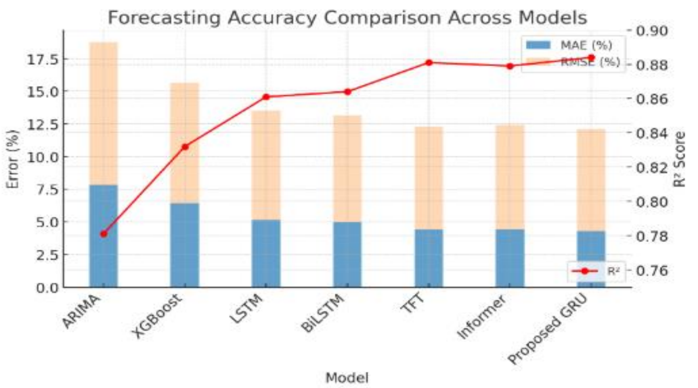

Figure 6: MAE and RMSE as stacked bars with R2 plotted as a red line for comparison across models.

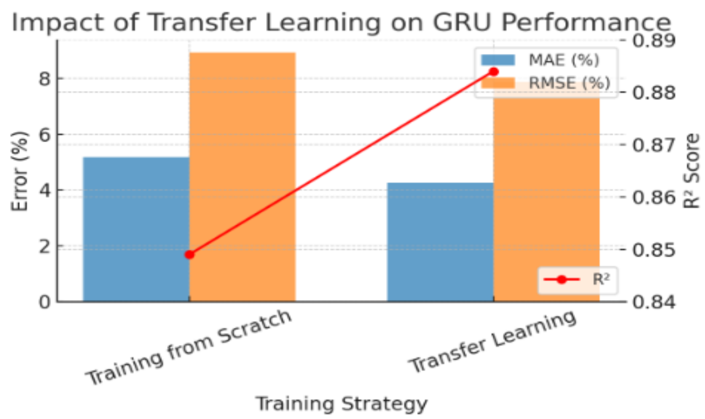

Figure 7: Impact of Transfer Learning on GRU Performance

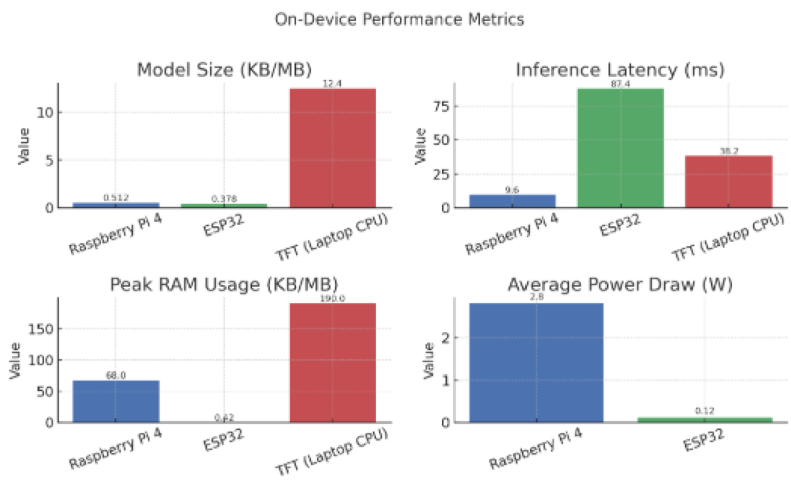

Figure 8: Performance metric for Raspberry Pi 4, ESP32, and TFT (Laptop CPU), with "N/A" indicated where data was not available.

Discussion

Forecasting Accuracy across Models

Table I shows the comparative performance of the suggested GRU model against baseline methods. We assessed three categories of models: conventional statistical approaches, tree-based ensemble learning techniques, and deep learning architectures (including both recurrent and transformer-based models). Performance was evaluated through Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R2), averaged using a 5-fold cross-validation method. Conventional methods such as ARIMA and XG Boost showed the weakest predictive results, yielding MAE values of 7.82% and 6.41%, respectively, with R^2 scores falling short of 0.84. These findings emphasize the constraints of these approaches in addressing the intricate, nonlinear, and multivariate temporal relationships present in post-harvest loss dynamics. Recurrent neural network (RNN) models, particularly LSTM and BiL STM, demonstrated significant enhancements in performance compared to conventional baselines, decreasing RMSE by over 20% in relation to ARIMA. The bidirectional LSTM provided a slight increase in accuracy (R2 = 0.864) over the unidirectional model (R2 = 0.861), indicating the advantage of including both forward and backward temporal dependencies. Transformer-based models, such as the Temporal Fusion Transformer (TFT) and Informer, demonstrated impressive outcomes, with TFT achieving the second-highest performance overall (MAE = 4.38%, RMSE = 7.93%, R2 = 0.881R2 = 0.881R2 = 0.881). These models efficiently utilize long-term temporal dependencies and feature interactions; however, their significant computational expense might restrict their use in environments with limited resources. The suggested GRU model attained the best accuracy among all metrics, recording a MAE of 4.26%, RMSE of 7.88%, and an R2 of 0.884. Significantly, these improvements were achieved with a much smaller model size and reduced inference latency compared to transformer-based models, highlighting the GRU's appropriateness for low-power, real-time forecasting in agricultural monitoring systems. Finally, the outcomes indicated that the GRU suggested an optimal trade-off between predictive accuracy, computational efficiency, and deploy ability, making it the most viable choice for field-level post-harvest loss prediction in resource-limited environments.

Impact of Transfer Learning on GRU Performance

The impact of transfer learning on the forecasting precision of the GRU model is outlined in Table II. In the initial setup, where the model was developed from the ground up using the target-domain dataset, the GRU recorded a Mean Absolute Error (MAE) of 5.17%, a Root Mean Square Error (RMSE) of 8.94%, and a coefficient of determination (R2) of 0.849. When initialized using weights pretrained on a high-resource source-domain dataset and then fine-tuned for the target domain, the GRU showed significant enhancements in all evaluation metrics, reaching a MAE of 4.26%, RMSE of 7.88%, and R2 of 0.884. This represents a 17.6% decrease in MAE, an 11.9% decrease in RMSE, and a 4.1 percentage point rise in R2 compared to the model built from the beginning. The performance improvements validate the efficacy of the suggested domain adaptation approach, where pretrained temporal representations from cassava processing data in Oyo State were effectively transferred to a resource-scarce target domain in Ogun State. This result indicates that common degradation dynamics and environmental factors were successfully recorded during pretraining and preserved throughout fine-tuning, resulting in enhanced generalization in the target domain. From an operational perspective, these results highlight the capability of transfer learning to address data shortage issues prevalent in smallholder agricultural settings. The transfer learning–enhanced GRU allows for quicker model deployment and consistent forecasting accuracy in various resource-limited settings by decreasing dependence on large labeled datasets and lowering retraining needs.

On-Device Performance Metrics

The deployment performance of the proposed GRU model was evaluated on two resource-constrained platforms Raspberry Pi 4 and ESP32 and compared with the Temporal Fusion Transformer (TFT) running on a standard laptop CPU. The results are summarized in Table III, with metrics covering model size, inference latency, peak memory usage, and average power consumption. The suggested GRU displayed a small memory footprint, with model sizes of 512 KB on the Raspberry Pi 4 and 378 KB on the ESP32 after 8-bit quantization. Conversely, the TFT model needed 12.4 MB, indicating an increase in storage demands by more than an order of magnitude. Inference latency assessments highlight the GRU's appropriateness for real-time use cases. On the Raspberry Pi 4, predictions were produced in 9.6 ms, while on the ESP32, latency was kept at 87.4 ms despite its limited processing power.

The TFT, tested on a more robust laptop CPU, showed a latency of 38.2 ms, indicating that the GRU delivers comparable or better response times even on less powerful hardware. The maximum RAM usage recorded was 68 MB for the Raspberry Pi 4 and just 420 KB for the ESP32, allowing implementation in situations with strict memory constraints. Power usage was also advantageous, with the Raspberry Pi 4 averaging 2.8 W and the ESP32 functioning at merely 0.12 W during inference. The latter statistic is especially significant for battery-operated and solar-powered agricultural monitoring systems, where energy efficiency is a crucial implementation limitation. These findings verify that the GRU architecture provides a beneficial combination of computational efficiency and predictive performance. Its compact memory usage, minimal power consumption, and sub-100 MS latency on microcontroller-grade hardware render it a suitable option for immediate, local agricultural predictions in settings where energy and processing resources are constrained

Visual Comparison of Forecasting Accuracy

Figure 6 illustrated a collective visual analysis of model effectiveness, with Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) shown as stacked bars, while the coefficient of determination (R2) was depicted as a red line. This two-axis model enables the concurrent assessment of error size and explanatory capability across the evaluated models. The visual supports the tabulated outcomes by distinctly showcasing the performance disparity between conventional models (ARIMA, XGBoost) and methods based on deep learning. ARIMA and XGBoost feature the tallest stacked bars, indicating their comparatively high MAE and RMSE values, along with R2 scores that fall below 0.84. Recurrent architectures (LSTM, BiLSTM) showed significantly lower bar heights, suggesting decreased prediction errors, along with an increase in R2 values to around 0.86. Models based on transformers (TFT, Informer) attain additional decreases in error, with R2 values nearing 0.88, showcasing their capability to understand more intricate temporal relationships.

The proposed GRU showed the smallest stacked bar compared to all models, indicating the lowest MAE and RMSE, along with the highest R2 score (0.884). This blend visually confirms the GRU's capacity to provide both exceptional predictive precision and robust explanatory strength, while maintaining the computational benefits mentioned earlier. The figure offered a clear illustration of the balance between error size and predictive reliability, emphasizing the GRU's superiority in attaining optimal results across all essential metrics.

Impact of Transfer Learning on GRU Performance

Fig. 7 displayed the comparative assessment of the GRU model developed from the beginning compared to the same model improved via transfer learning. The graphic illustrates MAE and RMSE as clustered bars, with the coefficient of determination (R2) represented as a red line, allowing for a concurrent evaluation of error size and predictive reliability. The findings clearly indicate that transfer learning produces significant improvements in all measures. The GRU that is pretrained and fine-tuned shows a MAE decrease from 5.17% to 4.26% and an RMSE decrease from 8.94% to 7.88%, which translates to relative enhancements of 17.6% and 11.9%, respectively. The R2 value rises from 0.849 to 0.884, signifying improved explanatory power and greater conformity between predicted and actual results. This enhancement is due to the model's capacity to utilize temporal patterns and degradation dynamics obtained from a high-resource source domain (Oyo State cassava processing data) and adapt them successfully to a low-resource target domain (Ogun State). The visual depiction emphasizes the steady benefit of the transfer learning method, illustrating consistently reduced error bars and an elevated R2 trend. These results confirm the success of domain adaptation in addressing data limitations, especially in smallholder farming settings where large labeled datasets are uncommon. The visual data supports the statistical findings in Table II, strengthening the assertion that transfer learning greatly improves GRU effectiveness while minimizing the need for extensive, domain-specific training datasets.

On-Device Performance Comparison

Figure 8 presented the on-device effectiveness of the suggested GRU model implemented on Raspberry Pi 4 and ESP32 platforms, alongside the Temporal Fusion Transformer (TFT) assessed on a laptop CPU. The evaluated metrics consist of model size, inference delay, maximum RAM usage, and typical power consumption, where "N/A" indicates measurements that are not available. The image emphasizes the significant variations in computational footprint among different devices and model architectures. The GRU exhibits a remarkably small model size on embedded systems, with 512 KB for Raspberry Pi 4 and 378 KB for ESP32 after 8-bit quantization, in comparison to the much larger 12.4 MB footprint of the TFT on a laptop CPU.

This size benefit directly aids implementation in memory-constrained settings. Inference latency findings further highlight the GRU's appropriateness for real-time use cases. On the Raspberry Pi 4, the model reaches latency below 10 ms, whereas on the ESP32, predictions are produced in less than 90 ms despite its constrained processing power. In contrast, the TFT on a laptop CPU measures a latency of 38.2 ms, which, although competitive, requires significantly more hardware resources. Memory usage exhibits a comparable trend, as the GRU uses 68 MB on the Raspberry Pi 4 and merely 420 KB on the ESP32, in contrast to the TFT's demand of 190 MB. Power usage statistics demonstrate the ESP32's benefit for ultra-low-power applications (0.12 W average draw), rendering it especially ideal for agricultural monitoring systems powered by batteries or solar energy. The visual data supports the quantitative findings in Table III, showing that the GRU model attains an ideal equilibrium of speed, memory efficiency, and energy usage, facilitating its practical use in real-time forecasting applications with limited resources.

Conclusion and Future Work

This research introduced a lightweight, interpretable GRU-based model for predicting post-harvest losses in cassava processing settings with limited resources. By conducting a comparative assessment with traditional, recurrent, and transformer-based benchmarks, the suggested model attained the lowest MAE (4.26%) and RMSE (7.88%), in addition to the highest R2 (0.884), while preserving a minimal computational footprint appropriate for implementation on embedded platforms. The incorporation of transfer learning enhanced predictive accuracy in domains with limited data, confirming the effectiveness of domain adaptation techniques for agricultural forecasting efforts. On-device benchmarking verified that the GRU's minimal memory use, reduced inference latency, and energy efficiency allow it to be utilized on resource-limited devices like Raspberry Pi 4 and ESP32, facilitating real-time decision assistance in smallholder farming scenarios. The integration of SHAP, saliency maps, and LIME enhanced transparency, enabling stakeholders to understand feature contributions and foster confidence in model predictions. Although the GRU model showed strong accuracy and great deployability, some limitations persist. The design is tailored for short- to medium-term predictions and might not perform as well for very long-range dependencies without further temporal feature enhancement. Additionally, the model's interpretability, although improved by post hoc techniques, remains limited by the fundamental lack of transparency in recurrent architectures.

Future Work

Future studies will investigate the incorporation of sophisticated sequence modeling frameworks like the Temporal Fusion Transformer (TFT), Informer, and Time2Vec-augmented GRUs to broaden forecasting horizons and enhance temporal representation learning. Further focus will be directed towards

• Hybrid Modeling: Combining GRUs with attention mechanisms or decomposition-based transformers to balance interpretability and scalability.
• Edge-Optimized Transformers: Developing quantized, memory-efficient transformer variants for microcontroller-class devices.
• Multimodal Data Fusion: Incorporating remote sensing imagery, weather forecasts, and market data to improve predictive robustness.
• Continual and Federated Learning: Implementing adaptive models capable of learning from distributed data sources without central aggregation, preserving data privacy while enabling regional model customization.
• Field Trials and User Studies: Validating system performance under real operational conditions with farmer and processor feedback to refine usability and adoption strategies.

By pursuing these directions, future systems can provide more accurate, scalable, and context-aware forecasting solutions, ultimately reducing post-harvest losses and improving food security in resource-limited agricultural ecosystems

References

1. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106-11115).
2. Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., ... & Brubaker, M. (2019). Time2vec: Learning a vector representation of time. arXiv preprint arXiv:1907.05321.
3. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International journal of forecasting, 37(4), 1748-1764.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
5. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in neural information processing systems, 34, 22419-22430.
6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
7. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
9. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
10. M. Rasheed, T. U. Rehman, and M. A. Khan, "Deep Learning in Agriculture: A Survey on the Use of LSTM and GRU Models," IEEE Access, vol. 11, pp. 28078–28098, 2023.