A Review of Attention-Enhanced GRU Models with STL Decomposition for Food Loss Forecasting

Ru Poh Tan[1]*, Siew Mooi Lim[2]*, Kuan Yew Leong[3], Shee Chia Lee[4], Siaw Hong Liew[5], Jun Kit Chaw[6]
Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[1, 2, 4]
Research and Development Department, A.I. System Research Co., Ltd., Kyoto, Japan[3]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia[5]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[6]

***Abstract*—Forecasting food loss with high accuracy is crucial for improving global food security, optimising supply chains, and supporting sustainability goals. However, conventional time series models and standard deep learning techniques, including recurrent neural networks (RNNs), often fall short in handling the irregularity, seasonality, and complexity inherent in food loss data. While Gated Recurrent Units (GRUs) offer advantages over traditional RNNs, such as mitigating vanishing gradients, they still face limitations in modelling long-range dependencies and noisy sequences. This paper reviews recent advancements aimed at overcoming these challenges by enhancing GRU-based models with attention mechanisms and seasonal-trend decomposition using Loess (STL). Evidence from related domains shows that attention mechanisms improve the capture of long-term dependencies and interpretability, while STL decomposition strengthens stability and accuracy by isolating seasonal and trend components. Hybrid GRU models that combine both approaches consistently outperform standalone methods, highlighting their promise for robust and interpretable forecasting. Though underexplored in the context of food loss, this paper identifies the research gap and advocates for domain-specific GRU–attention–STL architectures, offering a foundation for future empirical work to enable timely interventions and foster resilient, data-driven food systems.**

***Keywords*—GRU; food loss forecasting; attention mechanism; seasonal decomposition; STL; loess; time series; deep learning**

I. Introduction

This study tackles the pressing challenge of enhancing GRU-based forecasting for food loss prediction, a task that demands accurate modelling of complex, irregular, and temporally dependent data. Food loss, defined as the measurable reduction in food quantity or quality throughout the supply chain, is not only an economic inefficiency but also a social and environmental concern. Globally, approximately one-third of all food produced—about 1.3 billion tonnes—is lost or wasted every year, representing an estimated USD 1 trillion in direct economic losses and contributing nearly 8–10% of total greenhouse gas emissions [1]. Moreover, about 14% of food is lost between harvest and retail, 32% on food service and more, intensifying the strain on global food systems [2]. These staggering figures highlight the urgency of developing robust forecasting models that can enable timely interventions to minimise loss, improve efficiency, and promote sustainability.

Traditional statistical and deep learning methods have been applied to time series forecasting in various domains, yet they often struggle with the irregularity, seasonality, and noise inherent in food loss data. The shortcomings highlight the need for innovative architectures capable of capturing both fine-grained temporal dynamics and broad structural patterns within food loss datasets, as food loss poses serious implications for food security, economic sustainability [3], and environmental health, necessitating precise, interpretable forecasting models to guide data-driven interventions.

Recent advances in attention mechanisms and decomposition techniques offer promising solutions. Attention mechanisms, such as self-attention, temporal attention, and multi-head attention, dynamically prioritise informative time steps and enhance the interpretability of deep learning models by revealing which patterns contribute most to predictions. At the same time, Seasonal and Trend decomposition using Loess (STL) has proven effective in isolating trend and seasonal components from noisy signals, improving data quality and learning efficiency. However, although advances in deep learning [19], especially RNN-GRU and attention mechanisms, have shown strong potential in time series analysis across various domains [4], the integration of both seasonal-trend decomposition using Loess (STL) and attention in GRU architectures (GRU-attention-STL) for food loss forecasting remain underexplored.

The rationale for this new initiative is therefore twofold: first, to synthesise existing advancements in attention-enhanced GRUs and STL-based decomposition to demonstrate their relevance to food loss forecasting; and second, to advocate for the development of hybrid GRU–attention–STL architectures tailored to the unique challenges of food loss data. Such models are expected to not only boost forecasting accuracy and robustness but also provide interpretable insights into the underlying seasonal and temporal patterns, thereby supporting data-driven decisions in agriculture and supply chain management.

This paper is organised to provide a comprehensive review of recent advancements in GRU-based time series forecasting, with a specific focus on applications relevant to food loss prediction. Section II.A introduces the foundational concepts of GRU models and their role in time series modelling, outlining their structure, strengths, and limitations, particularly in

*Corresponding authors

capturing long-term dependencies and handling complex sequential data. Section II.B explores how attention mechanisms, such as Self-Attention, Bahdanau, Multi-head, Cross-Attention, Dual-Attention, Temporal-Attention and more, have been integrated with GRUs to dynamically prioritise relevant time steps and improve forecasting accuracy in various domains. Section II.C reviews the use of seasonal-trend decomposition using Loess (STL) to preprocess time series data, enabling GRUs to better learn from trend and seasonal patterns by reducing noise and structural complexity. Section III presents the methodological review framework, summarising and comparing existing studies in structured tables based on their use of attention mechanisms and seasonal-trend decomposition using Loess (STL) decomposition, while identifying design patterns, model architectures, and performance outcomes. Section IV discusses the observed relationships across these studies, highlighting how attention mechanisms enhance temporal adaptability and how seasonal-trend decomposition using Loess (STL) supports signal clarity, as well as the synergistic benefits of combining both techniques. Section V outlines key research gaps, emphasising the absence of GRU-attention-STL hybrids in food loss forecasting and posing future research questions around fusion strategies, generalisability, and dataset availability. Finally, Section VI concludes by reaffirming the value of these enhancements for developing more accurate, interpretable, and robust GRU-based forecasting models and encourages future empirical exploration within the food supply chain domain.

II. Literature Review

This section provides an overview of the existing literature related to GRU-based time series forecasting, with a focus on techniques that enhance model performance through attention mechanisms and time series decomposition. These methods have been widely studied and applied across various domains. Their combined application remains underexplored in the context of food loss forecasting. By reviewing key advancements and their outcomes, this literature review aims to highlight the strengths, limitations and potential of these techniques as foundational components for future food loss prediction models.

A. Gated Recurrent Unit in Time Series Forecasting

The GRU [5] is a type of recurrent neural network (RNN) designed for processing sequential data such as time series [6]. While RNNs, Long Short-Term Memory (LSTM) networks [7], and GRUs all model sequential data, they differ in complexity and performance. Standard RNNs often suffer from the vanishing gradient problem, limiting their ability to capture long-term dependencies [8]. LSTMs address this limitation with memory cells and gated mechanisms: input, forget, and output gates that help manage information flow [9]. GRUs, introduced as a simpler alternative to LSTMs, use gating mechanisms as well but with a more compact structure, offering efficient retention of relevant temporal information.

GRUs use only two gates: the update gate, which controls how much past information is retained, and the reset gate, which determines how much past information is forgotten. This streamlined structure enables GRUs to selectively preserve relevant data and discard noise, leading to more efficient learning of sequence patterns. The operations within a GRU cell are defined by equations in (1):

$$z_t = \sigma\,(W_z\,[h_{t-1},\,x_t] + b_z)$$
$$r_t = \sigma\,(W_r\,[h_{t-1},\,x_t] + b_r)$$
$$\widehat{h_t} = tanh(W_{\hat{h}}\,[r_t \circ h_{t-1}, x_t] + b_{\hat{h}})$$
$$h_t = z_t \circ \widehat{h_t} + (1 - z_t) \circ h_{t-1} \quad (1)$$

where z_t and r_t are the update and reset gate vectors respectively at time step t. W's are the learnable weight matrices and the b's are the biases. x_t is the input vector and the h's are the state vectors. $[\cdot,\cdot]$ denotes the concatenation of two vectors into a longer vector. $\tilde{h}_t$ is the candidate output activation while h_t is the updated hidden state or can be said as the final hidden state. $\circ$ denotes element-wise multiplication. For time $t = 0$, the hidden state variable is set to $h_0 = 0$. The update-gate u and reset-gate r, control the flow of information from one time step to another, and thus they are able to capture long-term dependency.

GRUs often match the performance of LSTMs while using fewer parameters and requiring less training time. Their balance of simplicity and effectiveness makes them well-suited for time series forecasting, particularly when short- to medium-term dependencies are key. Fig. 1 illustrates the internal cell structures of RNN, LSTM, and GRU [6].

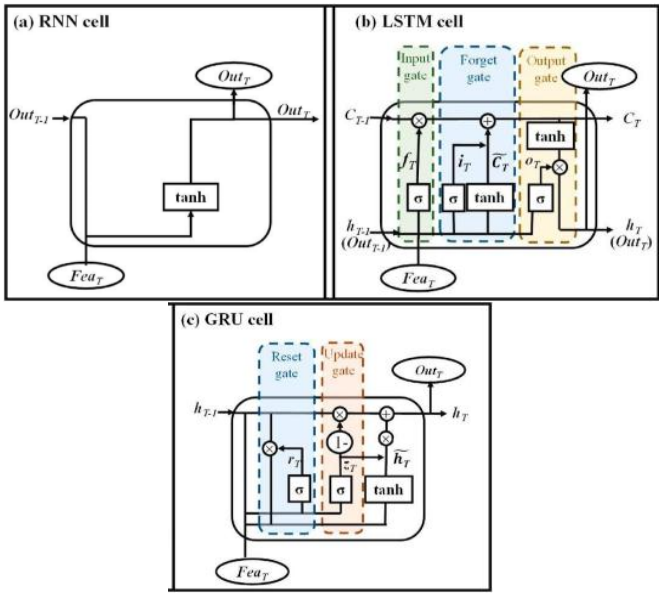


Fig. 1. Internal structure of (a) RNN cell; (b) LSTM cell; (c) GRU cell [6].

GRUs are widely used in time series forecasting for their ability to capture temporal dependencies. They are particularly effective in domains where patterns evolve gradually and depend on past values, such as weather prediction, energy demand, stock markets, and, increasingly, agriculture and food systems.

Fig. 2 shows the unfolded GRU through time [10]. $x_i^{(t)}$ is the input vector at time step vector at time step t. It represents the features or data being fed into the GRU at each time step. $h_i^{(t-1)}$, $h_i^{(t)}$, $h_i^{(t+1)}$, $h_i^{(t+2)}$, and, $h_i^{(t+n)}$ is the hidden state from the previous

time step. Each hidden state captures information from the current input and the previous hidden state. Each box labeled "GRU" processes input x and the previous hidden state h to produce the current hidden state $h_i^{(t)}$.

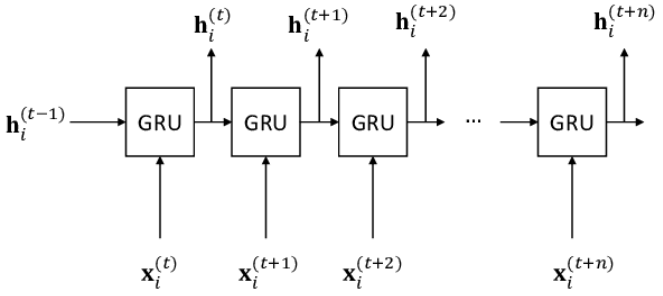


Fig. 2. GRU unfolded through time [10].

GRUs have been applied across various agriculture and food-related forecasting tasks, including crop yield prediction based on environmental data [11], soil moisture estimation [12], food demand forecasting in supply chain [13], spoilage prediction [14], seed trait analysis for protein, oil, and sucrose content [15], and price forecasting of agricultural products [14, 15]. Their ability to handle noisy, nonlinear, and seasonally influenced datasets makes GRUs a valuable tool in agricultural time series modelling.

With fewer parameters and faster training compared to LSTMs, GRUs maintain competitive accuracy [16, 17], particularly for capturing short- to medium-term dependencies, making them suitable for predicting food supply fluctuations and short-term losses. However, GRUs can struggle with highly irregular or seasonal data that require learning long-range dependencies [20]. To address this, recent studies have explored integrating attention mechanisms and time series decomposition to enhance GRU forecasting performance [21].

B. Enhancing GRU with Attention Mechanism

Attention mechanisms significantly enhance sequence learning by allowing models to dynamically focus on the most relevant parts of input sequences. Unlike standard recurrent models that rely on fixed-length context vectors, attention assigns varying weights to different time steps, improving performance on long and complex sequences.

1) Standard attention mechanism: The study [22] introduces a GRU-attention hybrid model for enhancing elastic scaling in container clouds. This model significantly improves prediction accuracy and resource management efficiency in container cloud environments. It effectively handles dynamic loads and reduces operating costs, although further optimisation is needed to control resource oversupply. Fig. 3 shows the prediction curve. Notice that GRU-attention appeared to be more accurate and less error than the GRU-only model.

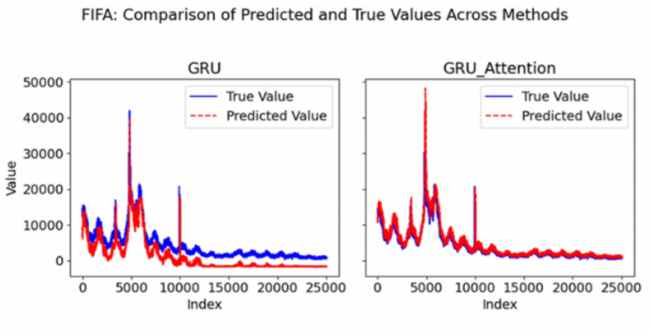


Fig. 3. Comparison of prediction accuracy [22].

To enhance multi-step solar radiation forecasting, Kong et al. [23] used the Kalman Filter [24] for data denoising, and improved input quality. They proposed a hybrid model combining Empirical Mode Decomposition (EMD), GRU, and standard attention mechanisms. Seasonal analysis showed this integration improved accuracy and reduced computational complexity. The structure of the attention-optimised GRU is shown in Fig. 4.

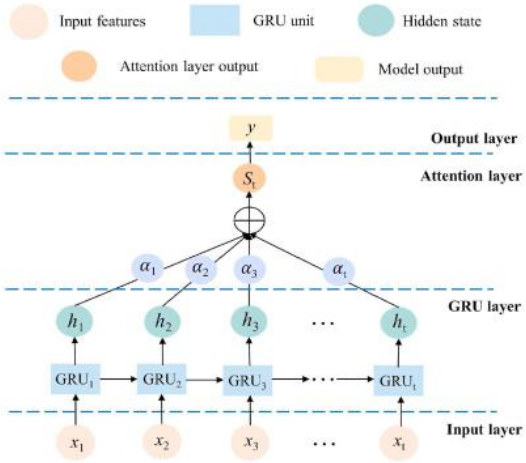


Fig. 4. Architecture diagram of an attention-optimised GRU [23].

2) Cross-attention and multi-scale attention: Cross-attention mechanisms vary in architecture, purpose, and integration. While [23] used standard cross-attention, Farahmand et al. (2025) demonstrate that cross-attention and multi-scale attention in the AttenGluco framework enhance blood glucose forecasting by effectively capturing long-term dependencies, leading to improved prediction accuracy [25]. Zhu et al. [26] also introduced MCI-GRU, a stock prediction model combining multi-head cross-attention with a modified GRU that replaces the reset gate with an attention mechanism to better capture temporal and cross-sectional features.

3) Self-attention and cross-attention: The general formula for the attention is presented in (2):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where Q, K, V are the concatenation of query, key and value vectors, and d_k is the dimension of the keys, used for scaling. The softmax makes the scores into probabilities that weight the

values V. In self-attention, the query, key and value all come from the same input sequence (3):

$$Q = XW^{Q}, \quad K = XW^{K}, \quad V = XW^{V} \tag{3}$$

where, X is the input sequence. Self-attention is used to let each position in the input sequence attend to other positions in the same sequence. Cross-attention is similar to self-attention. Self-attention focuses on relationships within a single input sequence, while cross-attention focuses on relationships between two different input sequences (4):

$$Q = X_1W^{Q}, \quad K = X_2W^{K}, \quad V = X_2W^{V} \tag{4}$$

where, X1 is the input where we want to generate attention for example decoder input, while X2 is the input to attend to, the encoder output. In short, the inputs for self-attention are all the same Q = K = V while cross-attention is from different sequences, Q ≠ K and V. Q comes from the decoder's output and K and V come from the encoder's output. The purpose of self-attention is to learn internal dependencies in a sequence. The cross-attention is to learn how one sequence attends to another.

In 2022, Shangzhu Jin et al. [27] enhanced deep knowledge tracing using a GRU network with self-attention, improving sensitivity to long sequences and prediction accuracy by leveraging historical knowledge states. The attention mechanism strengthened the influence of historical knowledge on future performance, leading to higher AUC and precision scores on the ASSISTments datasets compared to benchmark models. This demonstrated significant improvement in knowledge tracing tasks and better prediction performance. In the following year, Xikun Wei et al. [28] applied the Transformer's self-attention to improve runoff prediction in the Yangtze River basin. By integrating self-attention into LSTM and GRU models (LSTM-SA and GRU-SA), they achieved notable accuracy gains, especially under limited training data, by focusing on critical information.

The ProbSparse self-attention mechanism, a variant of traditional self-attention, was introduced to improve efficiency in long-sequence time series forecasting. Study [29] integrates it with LSTM to enhance prediction capacity for such tasks. Given the similarity between LSTM and GRU, integrating ProbSparse self-attention into GRU also holds potential.

Similarly, [30] showed that integrating multi-head self-attention into a GRU framework enhances the modelling of global dependencies and complex temporal patterns. Their FSR-MSAGRU model, applied to PM2.5 forecasting, improved accuracy and generalisation across datasets, reinforcing the value of attention in time-series prediction.

Demertzis et al. [31] showed that integrating GRUs with a cross-modal dynamic attention mechanism improves anomaly detection in time-series data. GRUs enhance prediction accuracy by capturing temporal dependencies and enabling adaptive focus on relevant features, particularly in smart communication environments.

4) Temporal attention and other attentions: Sehovac and Grolinger [32] demonstrated that temporal attention has improved time-series forecasting when combined with GRUs. Their work on electrical load prediction showed that incorporating Bahdanau attention into Seq2Seq RNNs enhances the capture of time dependencies, leading to higher accuracy than traditional models. Similarly, [33] integrated temporal and spatial attention into a GRU-based architecture, significantly boosting forecasting and fault diagnosis for electro-mechanical actuators. Their model, combining seasonal-trend decomposition using Loess (STL), hybrid spatial–temporal attention GRU, and DTW-based similarity, achieved accurate multistep predictions and robust fault detection, outperforming standard GRU and LSTM models. Fig. 5 illustrates the model's structure.

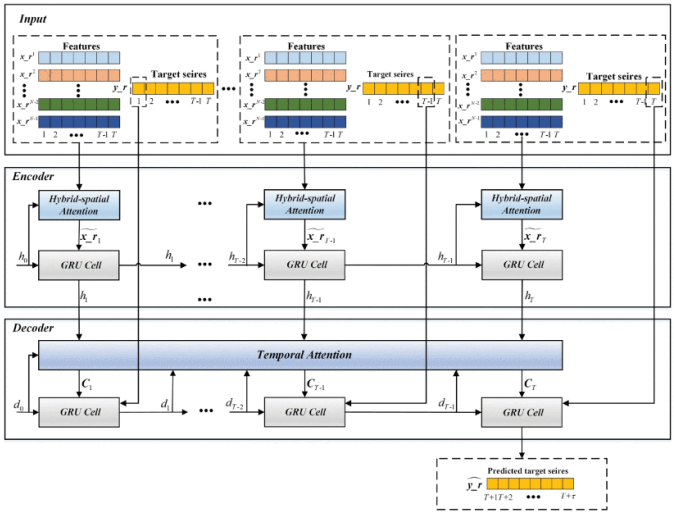


Fig. 5. Structure flowchart of HSTA-GRU model [33].

The authors in [34] investigated integrating GRUs with various attention mechanisms, including Multiplicative, Scaled Dot-Product, Bahdanau, Luong, and Self-Attention, to improve phishing URL detection. After comparing performance across these mechanisms, the study selected top-performing models and benchmarked them against GRU without attention and other baselines. As shown in Fig. 6, attention-enhanced models achieved higher accuracy and overall better performance, indicating that attention helps GRUs focus on relevant input features for more accurate threat detection.

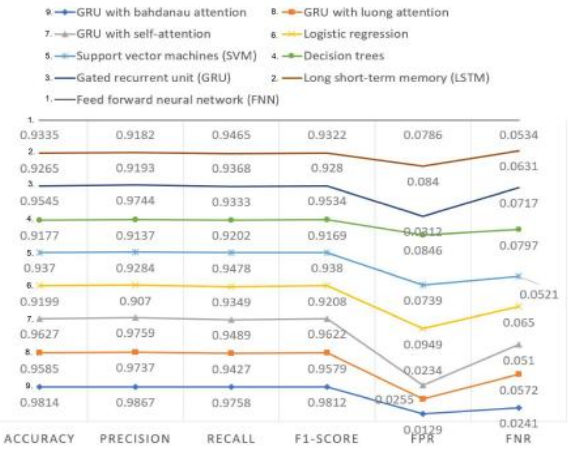


Fig. 6. Comparison of performances with various attentions and models.

The author in [35] introduces a GRU-based encoder-decoder model with a temporal attention mechanism for predicting ship speed over multiple horizons. The temporal attention mechanism and encoder-decoder structure enhance the model's ability to handle multi-horizon predictions and consider exogenous factors. By incorporating exogenous factors and focusing on influential time steps, the model achieves better prediction accuracy. Fig. 7 proves that by adding an attention mechanism to GRU, the average RMSE and average MAE prediction errors are reduced.

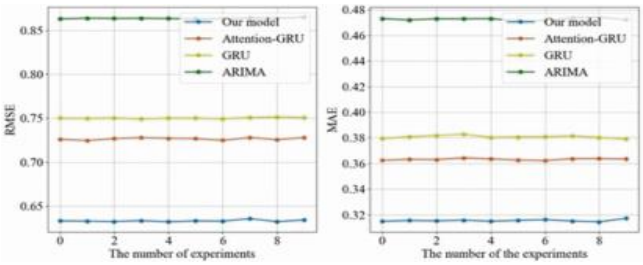


Fig. 7. The RMSE and MAE of different models of the ten experiments [35].

In [36], an interpretable condition monitoring method for offshore wind turbine gearboxes is proposed using a Spatial-Temporal Attention and GRU-based Network (STAGN). The model integrates spatial-temporal attention with GRU to effectively extract features from SCADA data. Fig. 8 shows the structure of the spatial attention module. Each input variable's spatial attention weight is calculated by (5) and (6) is used to calculate the normalised spatial attention weight αj(t) for the j-th input variable at time t.

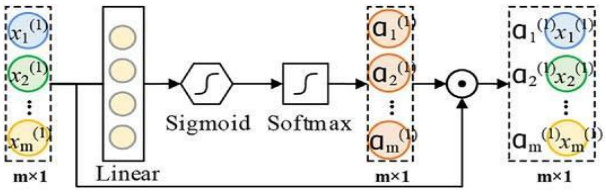


Fig. 8. Structure of spatial attention module [36].

$$e^{(t)} = \sigma(W_{sa}x^{(t)} + b_{sa}) \quad (5)$$

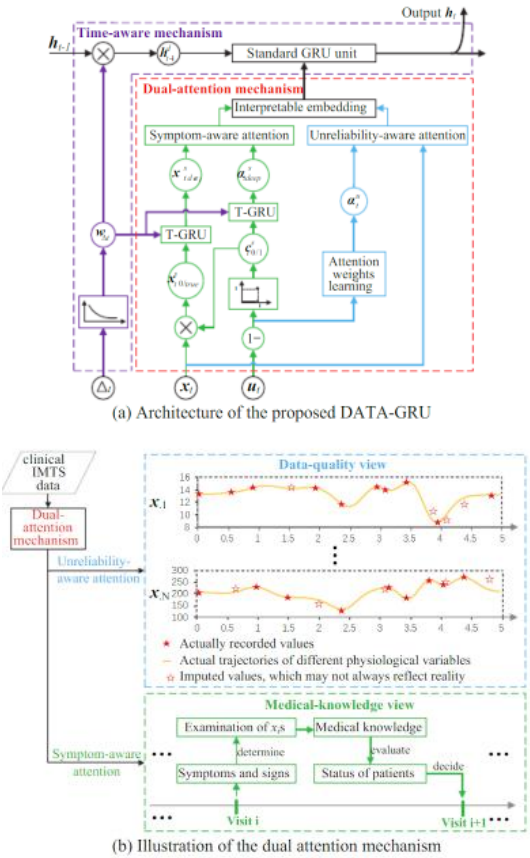
$$\alpha_j{}^{(t)} = \frac{exp(e_j{}^{(t)})}{\sum_{i=1}^{m} e_i{}^{(t)}} \quad (6)$$

According to [36], the temporal attention module enhances key temporal feature representation and mitigates early-stage information loss, while the spatial attention module captures spatial correlations. Together, they improve early fault detection and normal behaviour modelling (NBM), as validated on the Donghai Bridge offshore wind farm. Besides, Cui et al. [37] showed that integrating hierarchical contextual attention into GRU networks improves forecasting in sequential recommendation tasks by effectively capturing both long- and short-term dependencies. Their HCA-GRU model outperformed traditional RNN-based approaches in accuracy and ranking performance on real-world datasets.

5) Trend-aware attention: Trend-aware attention [38] is designed to highlight long-term upward or downward patterns in time-series data, unlike standard attention, which treats all inputs equally. Chen et al. [38] introduced the TAA-Transformer, which incorporates this mechanism to improve lithium-ion battery capacity prediction by capturing both local and global features. Compared to GRU, which lacks trend-aware attention, the TAA-Transformer achieved superior predictive accuracy. Although focused on Transformers, the study suggests that integrating trend-aware attention into GRU is feasible and could enhance performance in similar forecasting tasks.

Trend-aware multi-head flow attention has also been applied to forecasting tasks. Yang et al. [39] proposed JGFACN, a model combining a dynamic Jacobi graph with trend-aware flow attention for traffic prediction. Compared to methods including LSTM [18], JGFACN showed superior performance by dynamically adapting to temporal features. While the study did not integrate attention into GRU, its effectiveness suggests potential gains if applied to GRU-based time-series forecasting models.

6) Dual-attention: Tan et al. [40] also introduced DATA-GRU, a Dual-Attention Time-Aware GRU model, to address irregular intervals and missing values in EHR time-series data. Incorporating unreliability-aware and symptom-aware attention, the model effectively predicts patient mortality risk with improved accuracy and interpretability over state-of-the-art methods. Fig. 9 illustrates the DATA-GRU architecture and attention integration.



(a) Architecture of the proposed DATA-GRU

(b) Illustration of the dual attention mechanism

Fig. 9. Architecture of DATA-GRU is shown in (a) and dual attention mechanism in (b) [40].

The study in [41] proposed a dual-stage attention-based RNN (DA-RNN) using input attention to focus on relevant driving series and temporal attention to capture long-term dependencies, improving time-series forecasting accuracy on datasets like SML 2010 and NASDAQ 100. While based on LSTM, the dual-stage attention mechanism could similarly benefit GRU models. In 2023, [42] introduced PDAGRU, a Parallel GRU model with multiplicative and temporal attention layers, for probabilistic RUL prediction of wind turbine bearings. As shown in Fig. 10, this approach significantly enhances accuracy and reliability by better extracting degradation features from time-series data.

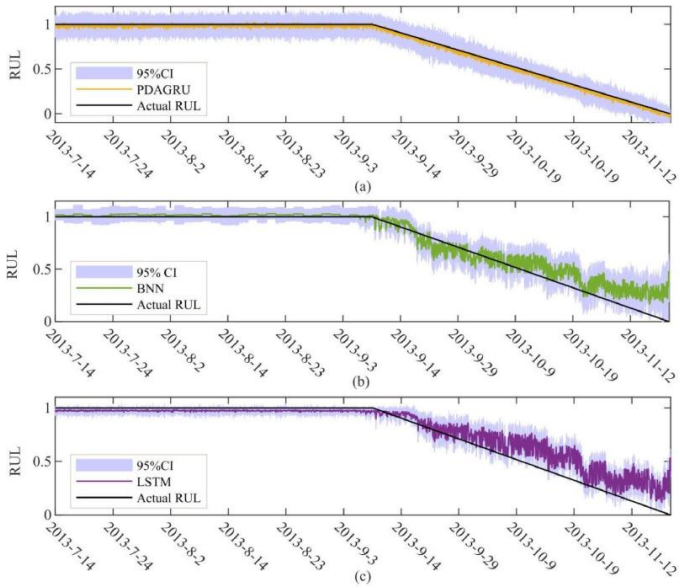


Fig. 10. The RUL prediction performance metrics with different methods [42].

In [43], the Hierarchical Attention Cascade Neural Network (CasHAN) is proposed to predict future cascade growth based on early evolution. As shown in Fig. 11, sequences from cascade graphs are encoded using a bidirectional GRU, then refined through node-level attention (capturing user influence) and sequence-level attention (addressing community redundancy), improving prediction accuracy.

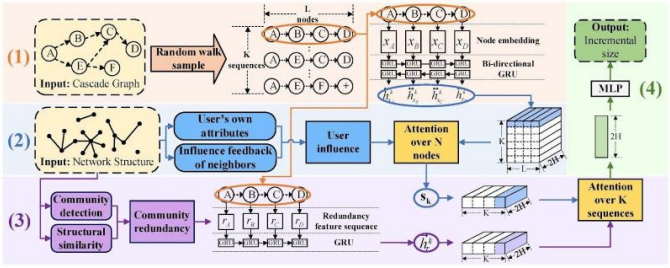


Fig. 11. The overall architecture of the proposed CasHAN model [43].

However, these attention-based GRU models have not been specifically applied to food loss forecasting, which poses distinct challenges due to the irregular, seasonal, and multi-source nature of agricultural data. Unlike applications in stock prediction, solar radiation, or aerospace forecasting, food loss prediction requires models to handle domain-specific variables, temporal variability, and data sparsity. This highlights a critical research gap for attention-enhanced GRU models tailored to food loss forecasting.

C. Enhancing GRU with Seasonal and Trend Decomposition

Seasonal-trend decomposition using Loess (STL) [44] is a time-series technique that separates data into seasonal, trend, and residual components. It uses LOESS, Locally Estimated Scatterplot Smoothing, and a nonparametric regression method, to flexibly estimate trends and seasonality over time. This decomposition aids in forecasting, anomaly detection, and identifying underlying patterns.

Despite advances in time-series modelling, decomposition integration within GRU architectures remains underexplored. Hewamalage et al. [45] noted that while RNNs, including GRUs, can model seasonality, this is mainly effective for time series with consistent seasonal patterns. For nonstationary or variable-seasonality data, they recommend deseasonalisation prior to modelling. Seasonal-trend decomposition using Loess (STL) provides a robust, nonparametric approach to decomposing time series into seasonal, trend, and residual components, enabling models to focus on learning more meaningful patterns from each.

Similarly, [46] proposed a short-term electricity forecasting model combining seasonal-trend decomposition using Loess (STL) with dual GRUs to separately capture global and local dependencies, significantly improving accuracy and robustness in handling nonlinear real-world data. Fig. 12 shows the structure of the proposed scheme.

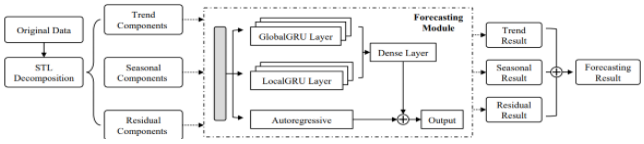


Fig. 12. The structure of the proposed scheme by [46].

The study in [47] introduced a hybrid STL-based framework for rainfall prediction, combining GRU, multi-time-scale GRU, and LightGBM to model decomposed components separately. This approach notably improved forecasting accuracy and robustness, especially for extreme rainfall events, by capturing key patterns in meteorological data. Fig. 13 illustrates the framework of the STL-ML approach to predict rainfall. As noted in [33], seasonal-trend decomposition using Loess (STL) also enhances attention-based GRU models by isolating trend, seasonal, and residual components, enabling more effective pattern extraction and improving fault prediction accuracy in electro-mechanical actuators.

An ensemble model proposed in 2020 [48] applied seasonal-trend decomposition using Loess (STL) with GRUs for base station traffic forecasting, demonstrating that modelling each decomposed component separately and recombining results enhances overall accuracy. Seasonal-trend decomposition using Loess (STL) effectively reduces noise and outliers, allowing GRUs to better capture nonlinear patterns and outperform standalone GRU and SARIMAX models.

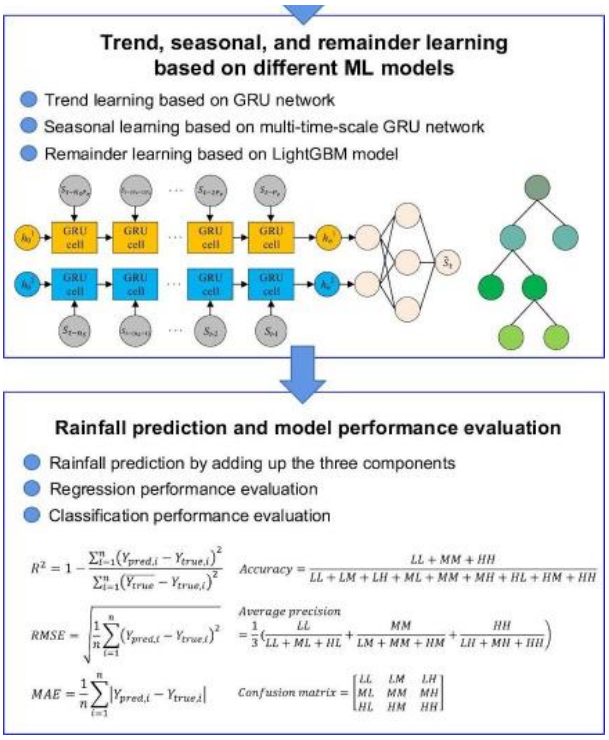


Fig. 13. Framework of the hybrid STL-ML approach for [47].

In climate forecasting, Liu et al. [49] introduced an STL-GRU model for temperature prediction, enabling GRUs to capture complex temporal correlations. This approach improved accuracy and efficiency across multiple regional datasets. The predictive performance indicators comparison for [49] are shown in Fig. 14. Similarly, Jia et al. [50] developed STL-IWOA-GRU for groundwater level prediction, integrating seasonal-trend decomposition using Loess (STL), an improved whale optimisation algorithm (IWOA), and GRU to handle data complexity with strong accuracy and robustness.

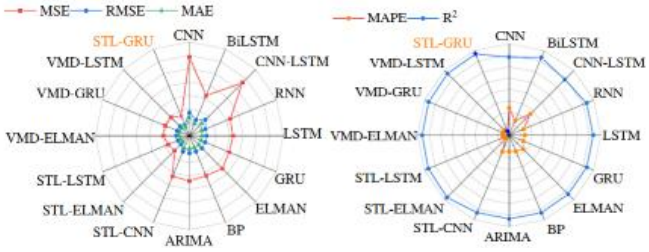


Fig. 14. Predictive performance indicators for [49].

Zhang et al. [51] proposed an STL-GRU-SVR model for sales forecasting, emphasising the importance of modelling trend, seasonal, and residual components independently to improve prediction. Likewise, [30] incorporated seasonal-trend decomposition using Loess (STL) in a GRU-based environmental forecasting model, enhancing feature space reconstruction and improving precision and robustness by isolating periodic and trend components.

III. Methodological Review Framework

This section presents a structured synthesis of existing studies that apply GRU-based models enhanced with attention mechanisms and time series decomposition. To make the findings accessible, the review is organised into a series of tables, each dedicated to a specific category of techniques. For instance, Tables I to VI summarise studies that integrate different types of attention mechanisms into GRU architectures, including Bahdanau (additive) attention, Luong (multiplicative) attention, self-attention, cross-attention, temporal attention, dual-attention, and more advanced designs such as hybrid spatial–temporal and trend-aware attention. Each table outlines the application domain, methodological approach, key contributions, and the specific challenges addressed, allowing us to see at a glance how these mechanisms influence forecasting performance across different contexts. Table VII focuses on studies that incorporate Seasonal-Trend decomposition using Loess (STL), highlighting how decomposing raw data into seasonal, trend, and residual components enhances learning efficiency and accuracy. By categorising and comparing these approaches, the framework provides both technical depth for researchers and clarity for practitioners, illustrating how attention mechanisms and seasonal-trend decomposition using Loess (STL) contribute individually and in combination to improving forecasting outcomes. Below shows review framework for each tables:

- Table I: Standard Attention Mechanisms
- Table II: Cross-Attention and Multi-Scale Attention
- Table III: Self-Attention and Cross-Attention
- Table IV: Temporal Attention and Other Attentions
- Table V: Dual-Attention Reviews
- Table VI: Trend-Aware Attention
- Table VII: Seasonal-Trend Decomposition using Loess (STL)

TABLE I. Review Framework for Standard Attention Mechanisms

Application	Methods	Key contributions	Gap Addressed
Container cloud elastic scaling [22]	GRU-attention hybrid model, Improved Horizontal Pod Autoscaler (HPA)	Enhanced load prediction and elastic scaling in container clouds, improving prediction accuracy and resource management efficiency.	Addressed the challenge of achieving efficient resource scheduling and elastic scaling in dynamic load environments, improving prediction accuracy and adaptability to complex load patterns.
Solar Radiation Prediction	EMD-GRU-Attention (EGA) model combining EMD, GRU, and **standard attention** mechanism.	Enhanced multi-step solar radiation prediction accuracy and reduced computational burden.	Addressed nonlinear, nonstationary data challenges and improved prediction accuracy.

TABLE II. REVIEW FRAMEWORK FOR GRU-CROSS-ATTENTION AND GRU-MULTI-SCALE ATTENTION MECHANISMS

Application	Methods	Key contributions	Gap Addressed
Blood Glucose Forecasting [25]	Multimodal Transformer with **cross-attention** and **multi-scale attention**.	Improved long-term BGL prediction accuracy using AI-READI dataset.	Addressed long-term forecasting challenges and data fusion with different rates.
Stock Prediction [26]	MCI-GRU model with **multi-head cross-attention** and improved GRU.	Enhanced stock prediction accuracy by integrating cross-attention and improved GRU.	Addressed limitations in capturing complex market dynamics and latent states.

TABLE III. REVIEW FRAMEWORK FOR GRU-SELF-ATTENTION AND GRU-CROSS-ATTENTION

Application	Methods	Key contributions	Gap Addressed
Knowledge Tracing [27]	GRU with **Self-Attention**.	Improved prediction accuracy and sensitivity to long sequence data in knowledge tracing.	Addressed information loss in long sequence data and enhanced historical knowledge impact on future performance.
Runoff Prediction in Yangtze River Basin [28]	Transformer (TSF), LSTM, GRU, LSTM-SA, GRU-SA with **Self-Attention**	Demonstrated that GRU outperforms TSF in limited data scenarios; Self-Attention improves LSTM and GRU performance.	Addressed the challenge of limited data for TSF and improved prediction accuracy with Self-Attention in LSTM and GRU.
Long Sequence Time-Series Forecasting [29]	Informer with **ProbSparse self-attention**, self-attention distilling, generative style decoder.	Achieved efficient long sequence forecasting with reduced time complexity and memory usage.	Addressed inefficiencies in Transformer models for long sequence forecasting.
PM2.5 Concentration Prediction [30]	FSR-MSAGRU model using feature space reconstruction and **multihead self-attention** GRU.	Improved prediction accuracy and generalisation by capturing periodic and global features.	Addressed limitations in capturing periodicity and global features in PM2.5 data.
Anomaly Detection in Data Streams [31]	CM-DANA model with **cross-modal dynamic attention** and multimodal learning.	Improved anomaly detection accuracy and efficiency in smart communication environments.	Addressed challenges in handling heterogeneous data types and evolving patterns.

TABLE IV. TEMPORAL ATTENTION AND OTHER ATTENTIONS

Application	Methods	Key contributions	Gap Addressed
Fault Diagnosis for Electro-Mechanical Actuators [33]	STL-HSTA-GRU with Dynamic Time Warping (DTW) for classification. HSTL = hybrid **spatial-temporal attention**	Improved fault diagnosis by combining STL decomposition with hybrid spatial-temporal attention GRU and DTW for classification.	Addressed challenges in handling nonlinear and seasonal data for early fault detection.
Load Forecasting [32]	Sequence-to-Sequence RNN with **Bahdanau Attention**, using GRU, LSTM, and Vanilla RNN cells.	Improved load forecasting accuracy by adapting S2S RNN with attention mechanisms.	Addressed challenges in capturing time dependencies and improving prediction accuracy.
Phishing URL Detection [34]	GRU with various attention mechanisms: **Multiplicative**, **Scaled Dot-Product**, **Bahdanau**, **Luong**, **Self-Attention**.	Achieved 98.14% accuracy using GRU with Bahdanau attention for phishing URL detection.	Enhanced detection accuracy and adaptability over traditional methods, addressing evolving phishing threats.
Ship speed prediction [35]	GRU-based encoder-decoder, **Temporal attention** mechanism, Dropout layers, Huber loss function, Adam optimizer	Proposed a model for accurate and timely ship speed prediction using a GRU-based encoder-decoder with a temporal attention mechanism.	Addressed the challenge of predicting ship speed by considering both previous speed and exogenous factors like weather and sea conditions, and improved prediction accuracy with a temporal attention mechanism.

Offshore Wind Turbine Gearbox Monitoring [36]	**Spatial-Temporal Attention** and Gated Recurrent Unit (GRU)	Introduced an interpretable model using spatial-temporal attention to enhance feature extraction and improve monitoring accuracy.	Addressed the challenge of early-stage information loss and lack of interpretability in condition monitoring models.
Sequential Recommendation [37]	**Hierarchical Contextual Attention**-based GRU (HCA-GRU)	Introduced a hierarchical contextual attention mechanism to enhance short-term interest modelling and combine it with long-term dependencies for better user interest representation.	Addressed the limitation of RNN's monotonic temporal dependency by capturing complex correlations among recent items and assigning nonmonotonic weights.

TABLE V. DUAL-ATTENTION REVIEWS

Application	Methods	Key contributions	Gap Addressed
Mortality Risk Prediction from Irregular Multivariate Time Series [40]	**Dual-Attention Time-Aware** Gated Recurrent Unit (DATA-GRU). The Dual-Attention including **Unreliability-aware attention** and **Symptom-aware attention**.	Introduced a model with time-aware and dual-attention mechanisms to handle irregular intervals and missing values in EHR data.	Addressed the challenges of irregular time intervals and missing values in EHR data by preserving temporal information and improving data reliability.
Time Series Prediction [41]	**Dual-Stage Attention**-Based Recurrent Neural Network (DA-RNN)	Introduced a dual-stage attention mechanism (input and temporal attention) to select relevant input features and capture long-term dependencies in time series.	Addressed the challenge of capturing long-term dependencies and selecting relevant driving series in time series prediction.
Probabilistic RUL Prediction of Wind Turbine Bearings [42]	Parallel GRU with **Dual-Stage Attention** Mechanism (PDAGRU). The Dual-Stage = **multiplicative attention** and **temporal attention**	Introduced a dual-stage attention mechanism to enhance degradation information extraction and a parallel structure for improved prediction accuracy and uncertainty quantification.	Addressed the challenges of efficient degradation information extraction, vanishing gradient problem, and prediction uncertainty.
Information Cascade Prediction [43]	**Hierarchical Attention** Cascade Neural Network (CasHAN) with **Node-Level** and **Sequence-Level Attention**	Introduced a model using hierarchical attention mechanisms to improve cascade prediction by considering user influence and community redundancy.	Addressed the lack of consideration for user influence and community redundancy in cascade prediction models.

TABLE VI. TREND-AWARE ATTENTION

Application	Methods	Key contributions	Gap Addressed
Lithium-Ion Battery Capacity Prediction [38]	**Trend-Aware Attention** (TAA) Transformer with 1-D Convolution.	Introduced TAA to capture local trends and global features, improving prediction accuracy over standard Transformers.	Addressed the limitation of standard Transformers in capturing local trend information in time-series data.
Traffic Forecasting [39]	Dynamic Jacobi Graph and **Trend-Aware Flow Attention** Convolutional Network (JGFACN)	Introduced a model combining dynamic Jacobi graph convolution with trend-aware flow attention to capture spatial-temporal correlations effectively.	Addressed challenges in capturing spatial-temporal correlations, dynamicity, and heterogeneity in traffic data.

Table VII below presents a review of studies that incorporate seasonal-trend decomposition using Loess (STL) as part of their forecasting frameworks. These works demonstrate how it is used to isolate trend and seasonal components from raw time series data, thereby improving model stability and learning efficiency. The table summarises how seasonal-trend decomposition using Loess (STL) has been combined with models such as GRU, LightGBM, and hybrid neural networks across various domains. This synthesis highlights the versatility

and effectiveness of seasonal-trend decomposition using Loess (STL) as a preprocessing technique, offering valuable guidance for its potential application in food loss forecasting.

TABLE VII. REVIEW FRAMEWORK FOR SEASONAL-TREND DECOMPOSITION USING LOESS

Application	Methods	Key contributions	Gap Addressed
Short-Term Electricity Forecasting [46]	Combined GRU with STL Decomposition	Developed a scheme using STL decomposition and dual GRU networks (GlobalGRU and LocalGRU) to capture local and global dependencies, enhancing prediction accuracy.	Addressed the inability of traditional models to capture complex nonlinear relationships and local historical information in electricity data.
Rainfall Time Series Prediction [47]	STL, GRU, Multi-time-scale GRU, LightGBM	Developed a hybrid STL-ML approach integrating STL decomposition with machine learning models to predict rainfall, effectively capturing trend, seasonal, and remainder components.	Addressed the challenge of accurately predicting rainfall by fully extracting and utilising underlying patterns and information from rainfall time series.
Fault Diagnosis for EMAs [33]	STL-HSTA-GRU, SM	Developed a method combining STL for seasonal-trend decomposition with HSTA-GRU for spatio-temporal prediction and SM for classification.	Addressed early-stage fault detection and nonlinear, seasonal data challenges in EMAs.
Base Station Traffic Forecasting [48]	STL, GRU	Developed an ensemble model combining STL decomposition with GRU to improve traffic forecasting accuracy by reducing noise and outliers	Addressed the challenge of nonlinearity and chaotic behaviour in traffic data, enhancing prediction accuracy over standalone models.
		.	
Climate Time Series Forecasting [49]	STL, GRU	Developed a hybrid STL-GRU model to improve temperature prediction accuracy by capturing seasonal variations and long-term trends.	Addressed the challenge of nonlinearity and temporal dependencies in temperature data, enhancing prediction accuracy over traditional models.
Groundwater Level Prediction [50]	STL, IWOA, GRU. IWOA stands for Improved Whale Optimisation Algorithm.	Developed a hybrid STL-IWOA-GRU model to improve GWL prediction accuracy by enhancing convergence speed and global search capabilities.	Addressed the challenge of nonlinearity and complexity in GWL time series under EWR, improving prediction accuracy over traditional models. EWR stands for Ecological Water Replenishment.
Sales Forecasting [51]	STL, SVR, GRU	Developed a sales forecasting model combining STL with SVR and GRU for higher accuracy.	Addressed the lack of direct forecasting using decomposed components in sales data.
PM2.5 Prediction [30]	FSR-MSAGRU (STL, MSAGRU)	Developed a model combining STL for capturing periodic and trend information with feature space reconstruction and multihead self-attention GRU for PM2.5 prediction.	Addressed the neglect of periodic and global features in PM2.5 prediction models.

IV. DISCUSSION

After reviewing a broad range of GRU-based forecasting studies that utilise attention mechanisms and seasonal-trend decomposition using Loess (STL), several clear relationships and patterns emerge across different domains and

methodological designs. These patterns not only clarify how these methods contribute to forecasting performance, but also offer valuable direction for future food loss forecasting research.

A. Attention Mechanisms

Various forms of attention, such as Self-attention, multi-head, cross-attention, dual-stage attention, and trend-aware attention, Bahdanau, Luong, and more, have been applied to GRU-based or related-based models like LSTM or others. A consistent pattern observed is that attention mechanisms improve the model's ability to capture long-term dependencies and enhance interpretability, especially in nonstationary or highly dynamic time series tasks. For instance, studies using multi-head or dual-stage attention (e.g. MCI-GRU [26], PDAGRU [42]) often outperform standard GRUs by dynamically focusing on critical time steps or features.

However, model complexity and training cost tend to increase with more advanced attention designs. These mechanisms also differ in scope while self-attention captures internal temporal dependencies, cross-attention and dual-stage attention are more effective when fusing heterogeneous inputs or signals from different domains.

B. Seasonal-Trend Decomposition Using Loess (STL)

Studies applying seasonal-trend decomposition using Loess (STL) consistently show that separating trend and seasonal components improves the stability and learning efficiency of GRU models. Models such as STL-HSTA-GRU [33], STL-IWOA-GRU [50] and more demonstrate that denoising inputs via seasonal-trend decomposition using Loess (STL) enhances model focus and leads to higher forecasting accuracy across domains such as electricity load, rainfall, climate, and groundwater levels. An important correlation observed is that seasonal-trend decomposition using Loess (STL) often pairs well with ensemble or hybrid frameworks, especially where trend clarity and seasonal structure play a strong role in outcomes.

C. Combination of Attention Mechanisms and STL

A smaller but growing subset of studies (e.g., STL-HSTA-GRU [33], FSR-MSAGRU [30]) explore hybrid models that integrate both attention mechanisms and seasonal-trend decomposition using Loess (STL). These models show synergistic benefits: seasonal-trend decomposition using Loess (STL) reduces input complexity by isolating components, allowing the attention-enhanced GRU to better identify which time points or signals are most important. This layered approach improves robustness and interpretability, particularly in domains with noisy, multi-scale, or highly seasonal data. While fewer in number, these hybrid designs consistently outperform single-technique models in accuracy and generalisation, suggesting that the combination of seasonal-trend decomposition using Loess (STL) and attention mechanisms is greater than the sum of its parts.

V. Research Gaps and Future Directions

A key limitation in current research is the lack of studies applying combined seasonal-trend decomposition using Loess (STL) and attention-based GRU architectures specifically to food loss forecasting. While this hybrid approach has proven effective in domains like electricity pricing, air quality monitoring, and fault detection, its potential remains unexplored in the context of food loss, despite the data's seasonal, irregular, and nonlinear nature.

Future work should focus on developing GRU-based hybrid models that incorporate both seasonal-trend decomposition using Loess (STL) and attention mechanisms. Such architectures could better capture long-term dependencies, dynamic temporal features, and structured patterns, offering a promising path toward improved accuracy and interpretability in food-related time-series forecasting.

A key direction for future research is optimising how attention weights are fused with decomposed components, trend, seasonality, and residuals, rather than applying attention to raw sequences. This component-level integration could enhance both interpretability and precision. Another open question involves the use of adaptive decomposition techniques that adjust to the characteristics of different food types or supply chain contexts, potentially improving generalisability across agricultural sectors. Additionally, the development and release of real-world food loss datasets and standardised benchmarks are critical for enabling rigorous model evaluation and comparison.

Addressing these gaps is essential for building practical, data-driven forecasting tools capable of making a tangible impact on food loss. By combining seasonal-trend decomposition using Loess (STL), GRU architectures, and attention mechanisms, future models can deliver more accurate, interpretable, and responsive predictions tailored to the complexities of food supply chains. Such advancements would support producers, distributors, and policymakers in identifying loss-prone points earlier, improving resource allocation, and enabling timely interventions. Ultimately, this research not only advances time-series modelling but also contributes to broader goals of reducing waste, enhancing food security, and fostering more resilient and sustainable supply systems.

VI. Conclusion

In summary, this paper highlights that both attention mechanisms and seasonal decomposition are promising enhancements for GRU-based food loss forecasting. Our review of models across domains, such as electricity load, air quality, rainfall, and fault detection, consistently shows performance gains when attention or decomposition techniques are used, especially in combination with GRU networks. Attention mechanisms improve a model's ability to capture long-range temporal dependencies and dynamically prioritise time steps, while seasonal-trend decomposition using Loess (STL) isolates trend and seasonal components, allowing GRUs to learn from cleaner, more structured inputs.

This study advances understanding of how architectural innovations can address challenges like long-term dependencies, seasonality, and noise in food loss data. It also identifies a key research gap: despite their success in other time-series contexts, hybrid STL-attention-GRU models remain largely unexplored in food supply chains. Given the inherent complexity and variability in food production and distribution,

developing such models is not only a valuable academic pursuit but also a practical necessity for reducing waste and improving food system resilience.

This review also sets the stage for future research by outlining key directions, such as optimising the integration of attention weights with decomposed components, exploring adaptive or domain-specific decomposition techniques for improved generalisation, and developing real-world datasets and benchmarks for consistent, reproducible evaluation.

Future empirical work should focus on designing and testing hybrid architectures that combine GRU networks with seasonal-trend decomposition using Loess (STL) and attention mechanisms for food loss forecasting. These models hold promise for delivering more accurate, interpretable, and resilient forecasting tools capable of supporting timely interventions and strategic decisions across the food supply chain. Ultimately, such innovations can contribute to global efforts to reduce food waste, strengthen supply chain resilience, and promote sustainable resource management.

ACKNOWLEDGMENT

This work is funded by Tunku Abdul Rahman University of Management and Technology (TAR UMT) through the Internal Research Grant: UC/I/G2025-00164.

REFERENCES

[1] C. Cederberg and U. Sonesson, Global food losses and food waste: extent, causes and prevention; study conducted for the International Congress Save Food! at Interpack 2011, [16 - 17 May], Düsseldorf, Germany. Rome: Food and Agriculture Organization of the United Nations, 2011.

[2] H. Forbes, T. Quested, and C. O'Connor, *Food Waste Index Report 2021*. Nairobi: United Nations Environment Programme, 2021.

[3] S. Negi, L.-S. Fan, H. Kim, T. Hidaka, A. Rani, and S.-Y. Pan, "Unlocking the potential of global greenhouse gas mitigation by reducing food loss and waste," *J. Agric. Food Res.*, vol. 21, p. 101925, Jun. 2025, doi: 10.1016/j.jafr.2025.101925.

[4] J. Rokui, "Historical time series prediction framework based on recurrent neural network using multivariate time series," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, Jul. 2021, pp. 486–489. doi: 10.1109/IIAI-AAI53430.2021.00084.

[5] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.

[6] X. Li, X. Ma, F. Xiao, C. Xiao, F. Wang, and S. Zhang, "Time-series production forecasting method based on the integration of Bidirectional Gated Recurrent Unit (Bi-GRU) network and Sparrow Search Algorithm (SSA)," *J. Pet. Sci. Eng.*, vol. 208, p. 109309, Jan. 2022, doi: 10.1016/j.petrol.2021.109309.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[8] S. Hochreiter, "Recurrent Neural Net Learning and Vanishing Gradient," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, 1998.

[9] S. Prakash, A. S. Jalal, and P. Pathak, "Forecasting COVID-19 Pandemic using Prophet, LSTM, hybrid GRU-LSTM, CNN-LSTM, Bi-LSTM and Stacked-LSTM for India," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, Mar. 2023, pp. 1–6. doi: 10.1109/ISCON57294.2023.10112065.

[10] M. Alfarraj and G. AlRegib, "Petrophysical-property estimation from seismic data using recurrent neural networks," in *SEG Technical Program Expanded Abstracts 2018*, Anaheim, California: Society of Exploration Geophysicists, Aug. 2018, pp. 2141–2146. doi: 10.1190/segam2018-2995752.1.

[11] A. P. Gopi, V. Swathi, G. S. Harshitha, B. Swetha, and N. Alekhya, "Prediction of Paddy Yield based on IoT Data using GRU Model in Lowland Coastal Regions," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Jan. 2023, pp. 1747–1752. doi: 10.1109/ICSSIT55814.2023.10060935.

[12] G. Wang, C. Wei, L. Yan, and J. Li, "Soil Moisture Prediction Model Based on Improved GRU Recurrent Neural Network," *Strateg. Plan. Energy Environ.*, Jan. 2024, doi: 10.13052/spee1048-5236.4329.

[13] K. Honjo, X. Zhou, and S. Shimizu, "CNN-GRU Based Deep Learning Model for Demand Forecast in Retail Industry," in *2022 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2022, pp. 1–8. doi: 10.1109/IJCNN55064.2022.9892599.

[14] Z. Guo *et al.*, "Multi-sensor fusion and deep learning for batch monitoring and real-time warning of apple spoilage," *Food Control*, vol. 172, p. 111174, Jun. 2025, doi: 10.1016/j.foodcont.2025.111174.

[15] Y. N. Kuan, K. M. Goh, and L. L. Lim, "Systematic review on machine learning and computer vision in precision agriculture: Applications, trends, and emerging techniques," *Eng. Appl. Artif. Intell.*, vol. 148, p. 110401, May 2025, doi: 10.1016/j.engappai.2025.110401.

[16] K. Kurumatani, "Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method," *SN Appl. Sci.*, vol. 2, no. 8, p. 1434, Jul. 2020, doi: 10.1007/s42452-020-03225-9.

[17] G. Avinash *et al.*, "Hidden Markov guided Deep Learning models for forecasting highly volatile agricultural commodity prices," *Appl. Soft Comput.*, vol. 158, p. 111557, Jun. 2024, doi: 10.1016/j.asoc.2024.111557.

[18] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," in *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, Shanghai, China: IEEE, Jun. 2020, pp. 98–101. doi: 10.1109/IWECAI50956.2020.00027.

[19] J. Koumar, T. Smoleň, K. Jeřábek, and T. Čejka, "Comparative Analysis of Deep Learning Models for Real-World ISP Network Traffic Forecasting," Mar. 20, 2025, *arXiv*: arXiv:2503.17410. doi: 10.48550/arXiv.2503.17410.

[20] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, "A review of irregular time series data handling with gated recurrent neural networks," *Neurocomputing*, vol. 441, pp. 161–178, Jun. 2021, doi: 10.1016/j.neucom.2021.02.046.

[21] G. Tuğba Önder, "Comparative time series analysis of SARIMA, LSTM, and GRU models for global SF6 emission management system," *J. Atmospheric Sol.-Terr. Phys.*, vol. 265, p. 106393, Dec. 2024, doi: 10.1016/j.jastp.2024.106393.

[22] Y. Zhang, Y. Sun, C. Song, and P. Gao, "A Container Cloud Elastic Scaling Method Based on GRU Attention Mechanism," in *2024 5th International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, Oct. 2024, pp. 290–293. doi: 10.1109/ICCEIC64099.2024.10775540.

[23] X. Kong, X. Du, G. Xue, and Z. Xu, "Multi-step short-term solar radiation prediction based on empirical mode decomposition and gated recurrent unit optimized via an attention mechanism," *Energy*, vol. 282, p. 128825, Nov. 2023, doi: 10.1016/j.energy.2023.128825.

[24] P. J. Hargrave, "A tutorial introduction to Kalman filtering," in *IEEE Colloquium on Kalman Filters: Introduction, Applications and Future Developments*, Feb. 1989, p. 1/1-1/6. Accessed: Jun. 08, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/197899.

[25] E. Farahmand *et al.*, "AttenGluco: Multimodal Transformer-Based Blood Glucose Forecasting on AI-READI Dataset," Feb. 14, 2025, *arXiv*: arXiv:2502.09919. doi: 10.48550/arXiv.2502.09919.

[26] P. Zhu *et al.*, "MCI-GRU: Stock prediction model based on multi-head cross-attention and improved GRU," *Neurocomputing*, vol. 638, p. 130168, Jul. 2025, doi: 10.1016/j.neucom.2025.130168.

[27] S. Jin *et al.*, "Self-attention based GRU neural network for deep knowledge tracing," in *2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, Dec. 2022, pp. 1436–1440. doi: 10.1109/ICIEA54703.2022.10006062.

[28] X. Wei, G. Wang, B. Schmalz, D. F. T. Hagan, and Z. Duan, "Evaluation

of Transformer model and Self-Attention mechanism in the Yangtze River basin runoff prediction," *J. Hydrol. Reg. Stud.*, vol. 47, p. 101438, Jun. 2023, doi: 10.1016/j.ejrh.2023.101438.

[29] H. Zhou *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, pp. 11106–11115, May 2021, doi: 10.1609/aaai.v35i12.17325.

[30] X. Yue *et al.*, "Novel hybrid data-driven modeling based on feature space reconstruction and multihead self-attention gated recurrent unit: applied to PM2.5 concentrations prediction," *Sci. Rep.*, vol. 15, no. 1, p. 17087, May 2025, doi: 10.1038/s41598-025-00911-9.

[31] K. Demertzis, K. Rantos, L. Magafas, and L. Iliadis, "A Cross-Modal Dynamic Attention Neural Architecture to Detect Anomalies in Data Streams from Smart Communication Environments," *Appl. Sci.*, vol. 13, no. 17, p. 9648, Aug. 2023, doi: 10.3390/app13179648.

[32] L. Sehovac and K. Grolinger, "Deep Learning for Load Forecasting: Sequence to Sequence Recurrent Neural Networks With Attention," *IEEE Access*, vol. 8, pp. 36411–36426, 2020, doi: 10.1109/ACCESS.2020.2975738.

[33] X. Zhang, L. Tang, and J. Chen, "Fault Diagnosis for Electro-Mechanical Actuators Based on STL-HSTA-GRU and SM," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–16, 2021, doi: 10.1109/TIM.2021.3127641.

[34] J. K S and A. B, "Exploring GRU-based approaches with attention mechanisms for accurate phishing URL detection," *Intell. Decis. Technol.*, vol. 18, no. 2, pp. 1029–1052, Jun. 2024, doi: 10.3233/IDT-240026.

[35] S. Dai and M. Yu, "Multi-Horizon Ship Speed Prediction with Temporal Attention Mechanism and GRU Encoder-Decoder," in *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, Mar. 2022, pp. 552–556. doi: 10.1109/CACML55074.2022.00099.

[36] X. Su, Y. Shan, C. Li, Y. Mi, Y. Fu, and Z. Dong, "Spatial-temporal attention and GRU based interpretable condition monitoring of offshore wind turbine gearboxes," *IET Renew. Power Gener.*, vol. 16, no. 2, pp. 402–415, Feb. 2022, doi: 10.1049/rpg2.12336.

[37] Q. Cui, S. Wu, Y. Huang, and L. Wang, "A hierarchical contextual attention-based network for sequential recommendation," *Neurocomputing*, vol. 358, pp. 141–149, Sep. 2019, doi: 10.1016/j.neucom.2019.04.073.

[38] C. Chen, Y. Wu, J. Shi, D. Yue, and H. Chen, "Leveraging Trend-Aware Attention in Transformers for Lithium-Ion Battery Capacity Prediction," *IEEE Sens. Lett.*, vol. 9, no. 6, pp. 1–4, Jun. 2025, doi: 10.1109/LSENS.2025.3562870.

[39] Y. Yang, Z. Yang, and Z. Yang, "Dynamic Jacobi graph and trend-aware flow attention convolutional network for traffic forecasting," *Digit. Signal Process.*, vol. 141, p. 104156, Sep. 2023, doi: 10.1016/j.dsp.2023.104156.

[40] Q. Tan *et al.*, "DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 01, pp. 930–937, Apr. 2020, doi: 10.1609/aaai.v34i01.5440.

[41] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 2627–2633. doi: 10.24963/ijcai.2017/366.

[42] L. Cao, H. Zhang, Z. Meng, and X. Wang, "A parallel GRU with dual-stage attention mechanism model integrating uncertainty quantification for probabilistic RUL prediction of wind turbine bearings," *Reliab. Eng. Syst. Saf.*, vol. 235, p. 109197, Jul. 2023, doi: 10.1016/j.ress.2023.109197.

[43] C. Zhong, F. Xiong, S. Pan, L. Wang, and X. Xiong, "Hierarchical attention neural network for information cascade prediction," *Inf. Sci.*, vol. 622, pp. 1109–1127, Apr. 2023, doi: 10.1016/j.ins.2022.11.163.

[44] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *J. Off. Stat.*, vol. 6, no. 1, pp. 3–73, 1990.

[45] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions," *Int. J. Forecast.*, vol. 37, no. 1, pp. 388–427, Jan. 2021, doi: 10.1016/j.ijforecast.2020.06.008.

[46] Y. J. Tian, S. J. Zhou, M. Wen, and J. G. Li, "A Short-Term Electricity Forecasting Scheme Based on Combined GRU Model with STL Decomposition," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 701, no. 1, p. 012008, Mar. 2021, doi: 10.1088/1755-1315/701/1/012008.

[47] R. He, L. Zhang, and A. W. Z. Chew, "Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning," *Knowl.-Based Syst.*, vol. 251, p. 109125, Sep. 2022, doi: 10.1016/j.knosys.2022.109125.

[48] K. Sebastian, H. Gao, and X. Xing, "Utilizing an Ensemble STL Decomposition and GRU Model for Base Station Traffic Forecasting," in *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, Sep. 2020, pp. 314–319. doi: 10.23919/SICE48898.2020.9240357.

[49] X. Liu and Q. Zhang, "Combining Seasonal and Trend Decomposition Using LOESS With a Gated Recurrent Unit for Climate Time Series Forecasting," *IEEE Access*, vol. 12, pp. 85275–85290, 2024, doi: 10.1109/ACCESS.2024.3415349.

[50] Z. Jia *et al.*, "A new strategy for groundwater level prediction using a hybrid deep learning model under Ecological Water Replenishment," *Environ. Sci. Pollut. Res.*, vol. 31, no. 16, pp. 23951–23967, Apr. 2024, doi: 10.1007/s11356-024-32330-0.

[51] C. Zhang and R. Shi, "Research on Sales Forecasting Model Based on GRU Neural Network and Machine Learning Model," in *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, Oct. 2023, pp. 575–579. doi: 10.1109/ICDSCA59871.2023.10392399.