# Memory Spike Detection System Using Gated Recurrent Unit and Transformer

**G. Lakshmi Supriya[1], K Mounika[2], Pearly Princess J[3]**

[1]Karunya University, School of Computer Science and Engineering, Karunya Nagar, Coimbatore, India
Email: *supriyayadav939[at]gmail.com*

[2]Karunya University, School of Computer Science and Engineering, Karunya Nagar, Coimbatore, India
Email: *kesireddygarimounika[at]karunya.edu.in*

[3]Karunya University, School of Computer Science and Engineering, Karunya Nagar, Coimbatore, India
Email: *pearlyprincess[at]karunya.edu.in*

**Abstract:** ***An intelligent deep-learning system called the Memory Spike Detection System was created to spot abrupt and unusual changes in memory utilization in computer environments. Conventional monitoring systems mostly rely on threshold-based warnings, which frequently miss small or quickly changing anomalies. Our suggested approach combines the advantages of Transformer and Gated Recurrent Units (GRU) architectures for effective temporal pattern learning and long-range dependency modeling in order to overcome this constraint. While the Transformer improves detection accuracy by modeling global attention throughout the memory usage timeline, GRU assists in capturing short-term sequential alterations at a lower computational cost. When combined, they allow for accurate forecasting and early detection of memory increases. This solution is appropriate for edge computing, servers, cloud platforms, and IoT devices for real-time monitoring. When contrasted to conventional machine-learning methods, experimental results demonstrate increased accuracy, fewer false alarms, and quicker anomaly identification. The suggested mixed paradigm shows great promise for preventing memory overload-related system failures and alert system management. Furthermore, the system's lightweight and scalable design makes it appropriate for deployment across edge computing devices, enterprise servers, cloud infrastructures, and IoT platforms that require real-time monitoring and rapid reaction. The suggested approach efficiently minimizes memory overload-related failures, increases system reliability, and enhances proactive alert management by allowing for early identification of anomalous memory behaviour. Overall, the hybrid GRU-Transformer framework is a promising and useful method for detecting intelligent memory anomalies in dynamic computing environments.***

**Keywords:** Gated Recurrent Unit (GRU), Transformer, Memory Spike Detection, Visualization, Time-Series Data , System Monitoring

**Abbreviations:**

| | |
|---|---|
| GRU | Gated Recurrent Unit |
| LSTM | Long Short- Term Memory |
| IOT | Internet Of Things |

## 1. Introduction

During the 1980s, when modern operating systems advanced and multitasking and time-sharing architectures became more widely used, researchers began to notice abnormal patterns in system memory behavior. Programs running on these systems, in particular, showed rapid and abnormal increases in RAM usage, which frequently appeared as short-term peaks during execution. These events, while not first formalized, were frequently known as memory utilization spikes. These surges were found to have a negative impact on system performance, causing slowdowns, resource contention, and, in some circumstances, system instability. As a result, analyzing memory usage spikes became an important component of performance evaluation, debugging, and operating system monitoring, laying the groundwork for future study in memory anomaly detection and resource management.

Advances in computer systems and Internet of Things (IoT) technology allow for intelligent healthcare applications, but resource and energy limits prevent effective depression detection in real-world IoT environments[1].This study compares the forecasting performance of RNN, LSTM, BiLSTM, GRU, and Transformer models on major global stock indices, showing that the Transformer consistently achieves higher precision and convergence efficiency across several assessment metrics[3].This paper discusses new AE-LSTM and AE-GRU encoder-decoder architectures for forecasting the values of various financial assets, proving that AE-GRU consistently outperforms AE-LSTM in capturing non-linear patterns and long-term dependencies under volatile market conditions [4].

The rapid spread of fake news on social media poses a serious societal threat, prompting this study to compare sequential and parallel memory-based deep learning models for text-based fake news detection, with transformer-based BERT outperforming across a variety of real-world datasets[7].This study examines advanced machine learning models for retail sales forecasting and finds that an optimized Random Forest model effectively captures complex seasonal patterns and multiple product-family interactions, outperforming traditional regression and other cutting-edge models [8].To address the challenges of existing statistical models in capturing the complex and non-linear dynamics of influenza outbreaks, this study presents a hybrid ARIMA-GRU framework that improves epidemic forecasting accuracy and promotes early public health intervention[9].

Paper ID: SR26217154244 DOI: https://dx.doi.org/10.21275/SR26217154244

Large volumes of information pertaining to performance are produced by modern computing systems, from embedded IoT devices to cloud servers. Among these, memory utilization patterns are crucial for preserving system performance and reliability. Unpredictable memory spikes can result in system slowdown, application crashes, or total system failure. These spikes can be brought on by malicious software, defective applications, memory leaks, or abrupt workload increases. Fixed threshold limits are often used in traditional memory monitoring techniques. However, because real-time workloads are extremely dynamic, non-linear, and unexpected, these approaches are frequently unreliable.

For modeling sequential and time-series system data, deep learning has become a potent tool. Specifically, Gated Recurrent Units (GRU) are appropriate for real-time settings because they efficiently learn sequential patterns with fewer parameters than LSTM. Transformer architectures, on the other hand, have transformed sequence modeling by introducing self-attention mechanisms that more successfully capture complicated global behavioral trends and long-range dependencies.

For reliable anomaly detection, this project suggests a Memory Spike Detection System that combines GRU and Transformer architectures. While the Transformer records broad patterns and abrupt abnormal shifts in memory usage, the hybrid model makes use of GRU's capacity to learn short-term and local temporal relations. This two-pronged strategy improves spike detection's precision and speed.

Data centers, enterprise servers, cloud platforms, multi-user applications, and resource-constrained IoT ecosystems are just a few of the settings where memory stability is critical. Compared to conventional threshold-based or classical machine learning techniques, the suggested method offers proactive detection, decreases downtime, and increases reliability by incorporating deep learning into system monitoring.

Spiking neural networks allow for energy-efficient, event-driven computation inspired by biological brains, but their practical deployment is hampered by hardware implementation of spike-based learning rules, highlighting the need for memristive devices in compact and reliable neuromorphic architectures[22].Beyond clinical localization, sEEG enables the study of large-scale neural dynamics and supports BCI systems, where understanding motor execution and imagery is key to improving control accuracy[23].

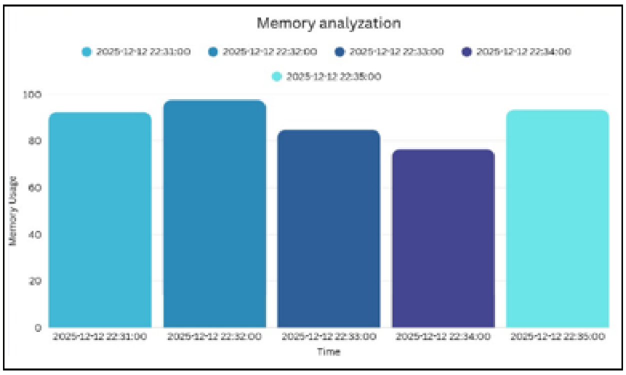

**Figure 1:** Analysis of the memory from the introduction

## 2. Theoretical underpinnings

This project combines computational and analytical approaches from diverse disciplinary areas such as time series analysis, deep learning, and intelligent system monitoring. The proposed framework relies on transdisciplinary ideas from machine learning, computer systems, and statistical anomaly detection, and incorporates methodological ideas such as predictive modeling, attention-based learning, and adaptive thresholding. These approaches combined can be used to develop a Memory Spike Detection System with the ability to detect sudden and irregular memory patterns in dynamic computing environments.

## 3. Objective

The main goal of this proposed work is to develop and deploy an intelligent deep learning-based Memory Spike Detection System that has the ability to properly model the temporal usage patterns of memory and detect any anomalies in memory spikes in dynamic computing systems. The proposed work intends to make use of a hybrid GRU-Transformer model that has the ability to properly capture both short-term and long-term dependencies in memory usage data, and also make use of adaptive error-based thresholding for proper anomaly detection. In addition to this, the proposed work also intends to minimize false positives and improve the accuracy of detection over traditional thresholding and machine learning approaches.

## 4. Literature Survey

The boundaries of data collection devices limit the use of objective depression assessments based on physiological data, which present new opportunities for clinical diagnostic support. Digital phenotypes of depression could be effectively captured and represented by countless digital devices, which are now deeply ingrained in daily life. Significant research and application efforts in this field have been hampered by fundamental issues with current research, such as device thresholds and inadequate use of distributed computational power. The feasibility of the suggested approach is validated by experimental results, which show that our model reduced inference power consumption by an average of 38% while achieving up to 70% correctness on the D-Vlog dataset [1].

Predicting financial market trends, particularly with regard to stock prices, is a topic that attracts a lot of interest and is very important. This is mostly because stock movements are constantly shifting and unpredictable. The events of the 2008 financial crisis serve as an example of how significant fluctuations can jeopardize the stability of global financial systems. Forecasting stock market trends has historically relied on techniques like technical and fundamental analysis. However, a growing need for more complex and sophisticated models emerges due to the unpredictable and turbulent environment marked by increased instability and the abundance of massive amounts of data influencing market patterns [2].

The volatility of crypto-currencies and other financial instruments makes it difficult to purchase and sell shares.

Numerous factors, including political, geographic, and socioeconomic factors, affect the dynamics of the stock and cryptocurrency markets. Changes in stock market trends are influenced by the significant variability among these various factors. Conventional techniques for forecasting stock prices include examining past data and trends to guide investment choices. These typically entail analyzing historical stock price data using statistical methods such as exponential smoothing, moving averages, and auto regressive integrated moving average (ARIMA) modeling. The temporal dependencies found in stock data, which are essential for precise stock market forecasting, are frequently ignored by these methods [3].

The emergence of social media platforms in the era of digitalization has completely changed how information is shared and consumed, allowing users to instantly exchange news, thoughts, and updates. Communication, information sharing, networking, marketing and advertising, entertainment, education, social activism, and more are all facilitated by it. In addition to the advantages of instant communication, social media has turned into a haven for the quick dissemination of false information, or "fake news." Because false information may affect public opinion, sway elections, and even incite violence, it poses a serious threat to public discourse, trust in institutions, and democratic processes. Misinformation, which is defined as inaccurate or deceptive information, is spreading online because to technical developments that make it simpler to alter images and videos. Sensational and deceptive posts frequently get more attention and interaction than later corrections, which is why this occurrence happens. By giving preference to content that receives a lot of interaction, algorithms on social media platforms contribute to the spread of false information by creating networks of persistent false information. Enhanced news and ideas spread quickly because these machines favor engagement over providing access to high-quality information [4].

Because of the dynamic operating conditions, inconsistent signals, and skewed class distributions, developing reliable diagnostic solutions for battery-powered systems continues to be a significant issue. These problems frequently plague traditional deep learning models, leading to decreased performance and flexibility. This study presents a thorough diagnostic methodology that combines adaptive hyper parameter optimization, spiking graph modeling, and multidimensional signal analysis in order to overcome these challenges. The methodology begins with an organized preprocessing stage that uses median and forward-fill techniques to deal with missing values, statistical filters to identify and remove outliers, and standardization to guarantee consistency parameter scaling. The Transform provides robust signal representation by capturing spectral and temporal characteristics. A Spiking Graph Transformer Network (SGTN), which combines attention-driven spatiotemporal learning with event-based neural computing, processes these representations after they are organized into graph sequences [5].

The problem of misleading information is not a novel idea. There is a need to reevaluate it because the rise of social networks and the internet in recent years has altered many of the concepts related to it. These days, news is disseminated more quickly and easily than ever thanks to social media's ever-growing appeal. In addition to harming social media sites, fake news can have negative effects on people and society in the real world. According to others, it weakens the human immune system, which speeds up the coronavirus's transmission. Regrettably, misinformation about COVID-19 has fueled the spread of disease and dysfunction among people, severely upsetting society. Additionally, false information is disseminated on social media platforms by artificial intelligence (AI) power bots, such as social bots or cyborgs, which automatically generate and spread fake material, broaden its audience, and give the impression of popular support. Computer algorithms that imitate human activity on social media are known as social bots. They automatically generate material, communicate with people, and disseminate malware, spam, rumors, incorrect information, slander, and even plain noise. Additionally, the widespread dissemination of false information has had a growing detrimental impact on business and stock markets [6].

There is a rising need for intelligent decision-support systems that can decipher complicated and unstructured feedback due to the massive volume of user-written product reviews produced by the quick growth of e-commerce platforms. Conventional recommendation techniques are still plagued by cold-start problems, poor interpretability, and an incapacity to capture long-range dependencies in lengthy evaluations, despite their widespread usage in sentiment analysis and rating prediction. In order to improve the precision and resilience of product suggestions, this study presents a hybrid Transformer–GRU model. While the GRU component effectively models sequential dependencies and changing user behavior over time, the Transformer module uses multi-head self-attention to extract deep contextual information from reviews. These findings demonstrate how well the Transformer–GRU hybrid architecture provides precise and customized e-commerce recommendations [7].

In order to reduce the impact of influenza epidemics on public health, early diagnosis and precise epidemics prediction are crucial since prompt action can successfully stop the disease's spread as well as the burden on healthcare systems. Although standard ARIMA models have demonstrated their value in short-term forecasting, especially in stable environments, they are less suited to handle rapidly developing epidemics due to their inability to keep up with the intricate and non-linear dynamics of disease dissemination. This is particularly true when outbreaks have complex seasonal patterns and erratic peaks that are difficult for ARIMA's to forecast on its own. [8].

An improved recurrent neural network architecture for binary categorization tasks, including analyzing sentiment on Amazon reviews, is presented in this article. The suggested model outperforms baseline models and provides robust generalization by combining GRUs, strategic dropout layers, L2 regularization in dense layers, and sophisticated optimization strategies. While feature significance assessments improve model interpretability, data balancing techniques provide equitable representation of both classes. Although performance has improved, issues with domain adaptation, computational resources, and managing unbalanced data still exist. In order to increase the applicability of RegGRUOpt across a variety of real-world text categorization scenarios, the future research will concentrate on investigating hybrid models, sophisticated regularization techniques, larger datasets, and improved accuracy [9].

Cyber dangers have also increased as a result of the quick rise in internet usage and connectivity. Zero-day attacks are among the most hazardous of these since they take advantage of weaknesses that system administrators are unaware of. Because they frequently rely on signature-based detection, traditional security solutions like firewalls and intrusion detection systems (IDS) are to identify new threats, particularly zero-day assaults. A transformer-based methodology for analyzing real-time network data and identifying zero-day risks is presented in this study. Transformers are a great option for network traffic analysis, where data arrives in the form of packets, because they are well known for their performance in NLP tasks and can accurately represent data in succession [10].

Enormous worldwide container hub terminals have enormous issues in their collection and distribution system as the amount of container marine traffic increases and the size of vessels increases. Furthermore, the peak of external truck arrivals worsens port congestion, resource waste, and air pollution emissions because most Chinese ports rely mostly on drayage trucks for the collecting and distribution of containers. By allocating truck arrival times and quantities, this approach greatly reduces traffic and resource waste during busy times at container terminals. In order to improve yard space allocation, crane setups, and the real time sequencing of container handling and pickup, container terminals increasingly use truck appointment data. This reduces needless container movements and increases yard operational efficiency [11].

Cyberbullying has become a major social issue in recent years due to the quick development of electronic technology and the extensive usage of the internet. Around 2000, a Canadian website owner who specialized in traditional bullying prevention first popularized the term "cyberbullying.This definition clarifies the technology component, the existence of animosity, the critical purpose to cause harm—a consensus among academics—and the recurring pattern of cyberbullying. Developing successful methods to lessen the negative consequences of cyberbullying on individuals and society requires an understanding of these basic characteristics [12].

Social media's rise has altered the communication landscape. It provides a forum for people to exchange stories and look for information on a range of topics, including general well-being. These platforms are online discussion boards where users can express concerns regarding asthma treatment, medication, triggers, and lifestyle modifications, as well as share personal experiences and look for support. Healthcare professionals and academics can obtain current data about public opinions, practices, and concerns around asthma management by comprehending the dynamics of these social media debates. However, multilingual complexities and cultural nuances present unique challenges when analyzing social media data, especially in languages other than English. One of the most widely spoken languages in the world is Arabic, which also includes conversations about asthma [13].

For real-time threshold-based sparse event identification, especially in extracellular neuronal spike recordings, online median estimate is essential. Due to their resilience to impulse noise, traditional sliding median and moving-window estimators have been employed extensively. However, they necessitate comparatively large buffers and auxiliary data structures, which increases estimator variance, hardware space, and power consumption. To increase speed and accuracy, a number of parallel and optimized median estimation methods have been put forth; nevertheless, these methods still rely on lengthy buffer lengths to produce acceptable variance and have trouble with quickly fluctuating, non-stationary data. The creation of median estimators that achieve lower variance with much smaller buffer sizes is motivated by recent research that highlight the need for compact, low-power, and fast-adapting estimators for implantable and real-time neural interfaces [14].

The development of high-performance brain–computer interfaces (BCIs) depends on the extraction of single-unit activity from intracortical recordings, which is essential to comprehending neural coding. Conventional spike detection and sorting techniques frequently struggle with noise, signal drift, and large-scale multichannel recordings because they rely on thresholding, feature extraction, and clustering. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have emerged as potent substitutes for automatic spike identification, classification, and prediction thanks to recent developments in deep learning. While hybrid CNN–RNN designs successfully capture both spatial and temporal brain dynamics, CNNs have shown excellent performance in learning discriminative spike features directly from raw signals [15].

Analyzing neuronal cooperativity requires spike sorting of simultaneously recorded extracellular signals, however this challenge is still difficult because of noise, spike overlap, and non-stationary recording settings. Traditional spike sorting algorithms, such as template matching and clustering-based approaches, sometimes struggle with overlapping spikes or signal drift and are unreliable in real time. Although a number of offline and semi-online methods have been put forth, the majority are not flexible or reliable in practical experimental settings. Therefore, online spike detection and classification techniques that improve signal-to-noise ratio and manage source separation have been the focus of recent study. Techniques like adaptive thresholding and linear filtering have demonstrated potential for increasing detection accuracy [16].

For large-scale brain interface, implantable neural recording devices need to have extremely compact, low-noise, and low-power recording channels. High power consumption and data bandwidth were the results of earlier neural front-end architectures that mainly concentrated on amplification and analog filtering, frequently sending raw data off-chip. In order to save power and communication overhead, recent research has moved toward combining on-chip digitization, spike detection, and feature extraction. To manage variability in neural signals and electrode settings, advanced CMOS-based neural recording channels now include calibration circuits, programmable gain amplifiers, and analog-to-digital converters. In order to sustain optimal gain and bandwidth over time, self-calibration techniques have been implemented. Ultra-low energy per conversion and low noise efficiency factors are becoming important performance indicators [17].

Low-latency, energy-efficient on-chip spike sorting solutions are in high demand due to the quick development of high-density neural recording systems, especially for closed-loop brain-computer interface applications. Due to frequent analog-to-digital conversions, conventional spike sorting systems significantly rely on digital signal processing, which increases latency and power consumption. To get around these restrictions, recent studies have looked into analog processing and in-memory computing. While neural network-based feature extraction offers greater versatility than handcrafted features, particularly for identical spike waveforms, delta-based spike representations have been demonstrated to increase noise robustness and decrease data redundancy. For compact and discriminative feature learning, autoencoder-based models have drawn interest [18].

Over the past 20 years, high-frequency oscillations (HFOs) and conventional epileptic spikes have gained recognition as accurate indicators of epileptogenic brain tissue. Automated detection techniques were developed since the manual identification of spikes, ripples, and ripples-on-spikes (RonS) in intracranial EEG is laborious and susceptible to inter-observer variability. Previous methods depended on manually created features and thresholding, which frequently lacked stability between patients and institutions. More precise and broadly applicable EEG analysis is now possible thanks to recent developments in artificial intelligence, especially deep learning. Temporal dependencies in EEG data can be effectively captured by recurrent neural networks, such as Long Short-Term Memory (LSTM) models [19].

Because they closely resemble biological neural processing and feature benefits like event-driven computation, online learning, and great energy efficiency, spiking neural networks (SNNs) have garnered a lot of attention. Nevertheless, it is still difficult to use traditional CMOS technology to implement biologically realistic learning rules like spike-timing-dependent plasticity (STDP) in hardware. Memristive devices' non-volatility, scalability, and analog weight storing characteristics have made them attractive candidates for artificial synapses, according to recent studies. Although there are still few large-scale implementations, early experiments have demonstrated the viability of memristor-based synapses in basic neuromorphic circuits. Basic brain capabilities like coincidence detection through STDP learning have been achieved in experimental implementations that combine memristive synapses with leaky integrate-and-fire neurons [20].

An essential preprocessing step in brain–computer interfaces (BCIs) and neural data analysis is the detection and categorization of action potentials from extracellular neural recordings. Spike identification and spike sorting are usually carried out as distinct, sequential processes in conventional methods, which may restrict performance in noisy recording environments. It is possible to identify and classify spikes more accurately by including waveform information straight into the detection step. In this regard, probabilistic models offer a rational framework for managing brain signal uncertainty. This paper presents a hidden Markov model-based method that models spike presence, neural identity, and waveform dynamics in order to jointly conduct spike detection and sorting. The suggested approach increases the dependability of BCI systems and improves neural decoding performance by generating probabilistic spike estimates instead of discrete counts [21].

Large-scale cloud infrastructure management and monitoring have become much more difficult due to the quick uptake of cloud computing. DevOps teams are burdened by traditional rule-based and threshold-driven monitoring solutions, which frequently fail to scale efficiently and necessitate significant manual intervention. Recent studies have investigated the application of deep learning and machine learning methods for automated anomaly detection in cloud systems in order to get beyond these restrictions. Neural network-based models have demonstrated a remarkable ability to identify anomalies in almost real time and capture intricate patterns across several system components. Previous research emphasizes how these methods can minimize service downtime, increase problem detection accuracy, and reduce operational overhead [22].

DevOps teams are burdened by traditional rule-based and threshold-driven monitoring solutions, which frequently fail to scale efficiently and necessitate significant manual intervention. Recent studies have investigated the application of deep learning and machine learning methods for automated anomaly detection in cloud systems in order to get beyond these restrictions. Neural network-based models have demonstrated a remarkable ability to identify anomalies in almost real time and capture intricate patterns across several system components. Previous research emphasizes how these methods can minimize service downtime, increase problem detection accuracy, and reduce operational overhead. These developments show how crucial sophisticated, automated monitoring systems are [23]

## 5. Implementation

**Block diagram:**

Proposed architecture for memory usage spike detection using parallel GRU and Transformer models, where preprocessed time-series data is analyzed to compute prediction errors, apply dynamic thresholding for anomaly detection, and visualize high-usage spikes through a dashboard and reporting interface.

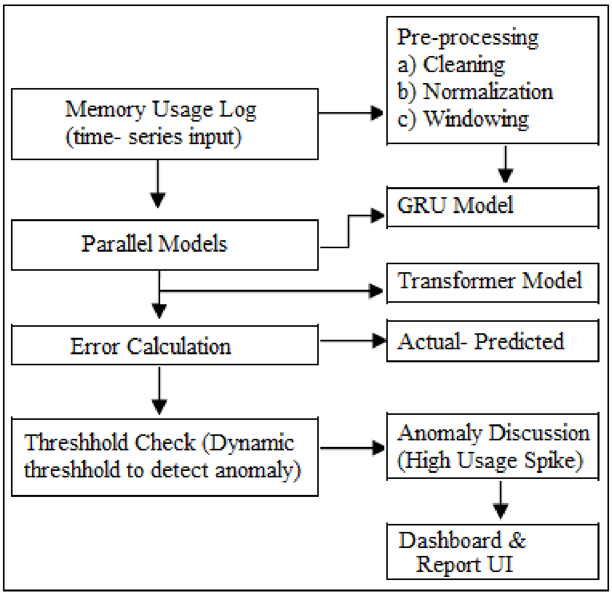

**Figure 2:** Memory spike detection system architecture

## 6. Methodology

The proposed Memory Spike detection System using gru and transformer methodology designed to identify abnormal memory usage patterns using these models. The system integrates Gated Recurrent Unit (GRU) network, Transformer Encoder, Anomaly detection based on statistical thresholds, a dashboard for seeing data and making reports.

### a) Data Processing

The dataset for this study consisted of memory consumption logs that were gathered from system monitoring tools. A timestamp and the associated memory use, represented as a percentage, make up each data sample. To enhance data quality and learning efficacy, a number of preparation procedures were used before model training. In order to lessen noise and short-term oscillations in the memory consumption signals, a moving average smoothing approach was first applied. To guarantee consistent feature scaling and quicker model convergence, the data were then normalized to the range [0,1] [0,1] using Min–Max scaling. A sliding window method with a sequence length of $T = 20$ T=20 was used for supervised learning, where each input sequence ⍰ ⍰ $= [mt, mt + 1, \ldots, mt + 19]$.The displays memory consumption numbers over 20 successive time steps, and the desired result is y=m t+20.The relates to the amount of memory used in the subsequent time step. In order to facilitate efficient model training, hyperparameter tuning, and objective performance evaluation, the processed dataset was finally split into training, validation, and testing sets in the proportions of 70%, 15%, and 15%, respectively.

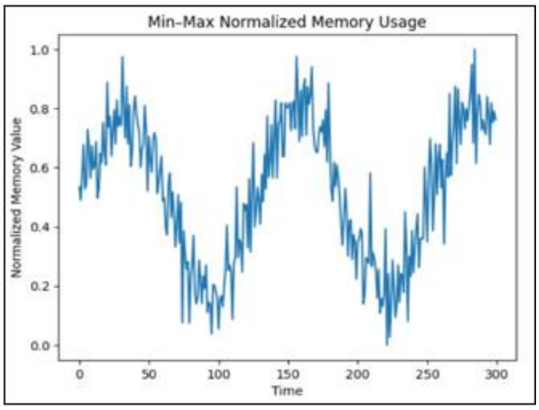

**Figure 3:** Time-Series Representation of Normalized Memory Usage

### b) GRU-Based Prediction Model:

Because of its effectiveness in capturing temporal dependencies in time-series data while retaining a comparatively low computing overhead, a Gated Recurrent Unit (GRU) architecture was chosen. A fully connected dense layer with 32 neurons with ReLU activation to improve non-linear feature representation comes after a GRU layer with 64 hidden units to learn sequential patterns in memory utilization. A single neuron in the final output layer predicts the subsequent memory use value. The Adam optimizer was utilized for effective and steady convergence during training, with Mean Squared Error (MSE) serving as the loss function. The model was trained using a batch size of 32 for 50 epochs. Compared to more complicated recurrent models, the GRU efficiently learns short-term and medium-term memory consumption trends thanks to its gated structure, which also requires less parameters and processing resources.

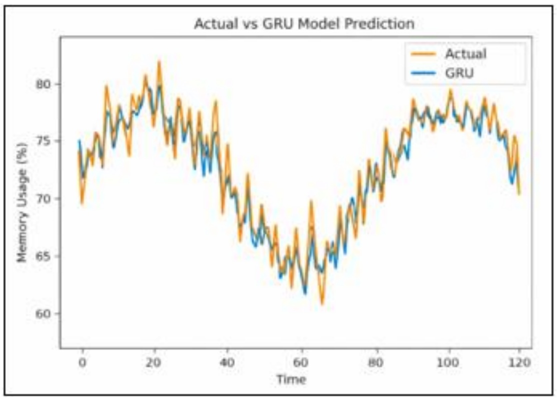

**Figure 4:** Min–Max Normalized System Memory Usage Over Time

### c) Transformer-Based Prediction Model:

Transformer models were used to identify seasonal trends and dependency relationships in memory utilization data. Transformers, in contrast to recurrent structures, rely on self-attention mechanisms that allow the framework to concurrently examine relationships throughout the whole input sequence. The Tra nsformer encoder, which consists of two stacked layers with four attention heads each, comes after positional encoding, which preserves sequential ordering data in the suggested architecture. The learnt models are further refined by a feed-forward network, and a final memory use estimate is produced by a dense output layer. The same training configuration—loss function, optimizer, number of epochs, and batch size—was used to guarantee an equitable and consistent comparison with the GRU-based model. Transformers are especially good at identifying abrupt spikes and intricate usage patterns in memory consumption because of their global attention capabilities; in these situations, they frequently beat conventional RNN-based methods.

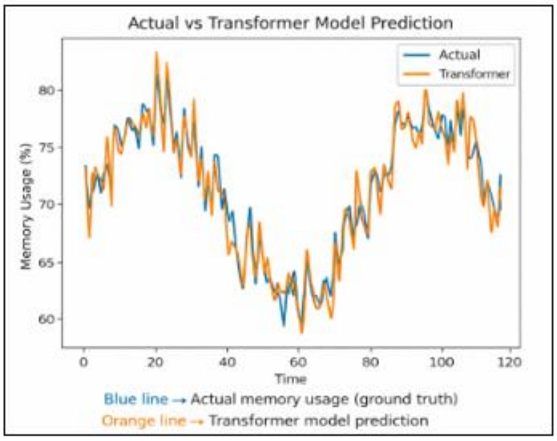

**Figure 5:** Comparison of Actual and Transformer-Based Memory Usage Predictions

**d) Anomaly Detection Strategy:**

Anomaly detection is carried out by calculating the prediction error at each time step following the generation of the expected memory use values using both the GRU and Transformer models. The definition of the absolute error is $Et =| mt - m\hat{\ } t |$, where mt shows how much memory is actually used, and m^t indicates the expected value at time t. Based on the statistical characteristics of the prediction mistakes, a dynamic threshold is computed to differentiate between abnormal patterns and normal behavior. In particular, the threshold is described as $Threshold = \mu E + k.\sigma E$, where $\mu E$ is the average of the prediction errors, $\sigma E$ is their standard deviation, and The sensitivity parameter k=2.5 was selected empirically. The relevant timestamp is marked as an anomaly, indicating strange or unexpected memory use behavior, if the prediction error at any time step surpasses this threshold ($Et$ > Threshold).

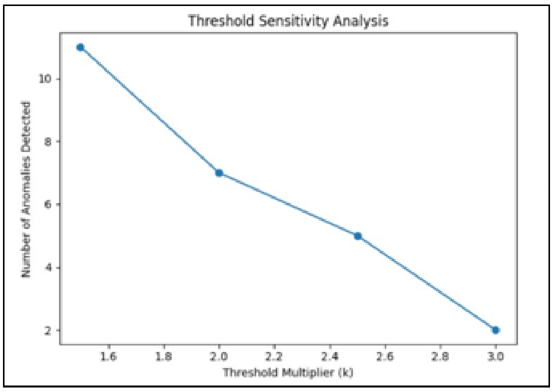

**Figure 6:** Threshold Sensivity Analysis

This illustrates Threshold Sensivity Analysis of the memory anomaly detection. As the threshold multiplier increases, the number of detected anomalies decreases, indicating reduced sensitivity.

## 7. Result analysis

### a) Memory Usage Behavior Analysis

When the memory consumption timeline is visualized, it shows a varying pattern with a steady increase in overall utilization. Abnormal memory behavior is indicated by a number of sharp spikes that surpass the dynamically calculated threshold. According to statistical analysis, peak memory consumption is between 98 and 99 percent, while average memory usage stays between 88 and 89 percent. Near these top utilization levels, several anomalous events are found, indicating times of memory saturation that could be brought on by resource-intensive apps or memory leaks.

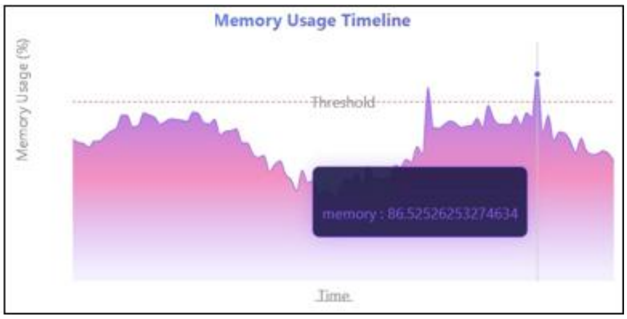

**Figure 7:** Memory Usage timeline

### b) Anomaly Detection Results

The dashboard's detected anomalies section shows that the anomaly detection module has successfully identified several anomalous events. The severity of these anomalies is divided into three categories: moderate high-usage cases (~79–85%), medium-severity rapid change occurrences (~98%), and critical high-usage events (~95–99%). By ensuring that only statistic

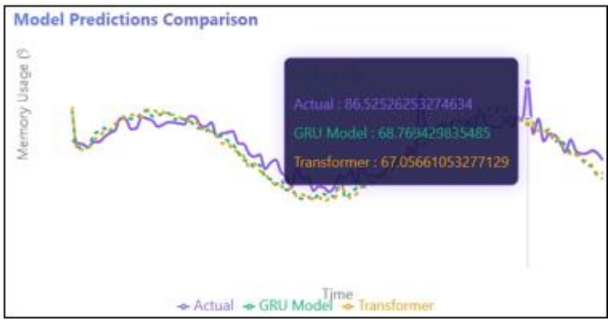

**Figure 8:** Models Prediction Comparsion

### c) Overall System Performance

The combined GRU–Transformer framework improves anomaly detection reliability, according to the overall experimental results. While dynamic thresholding successfully strikes a compromise between detection accuracy and false alarm rates, transformer models continuously outperform GRU models during abrupt memory use increases. The system's capacity to identify high-risk memory usage scenarios is validated by its reliable and consistent performance, indicating that it is suitable for real-world deployment.

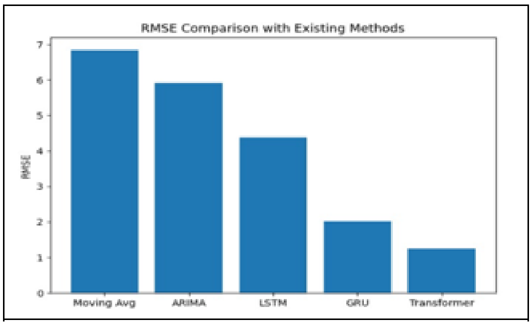

**Figure 9:** Comparison with existing models

This graph presents the RMSE comparison between the proposed deep learning models and existing prediction techniques. In this clearly GRU and Transformer has less error that means it is the best model to use in combine to get more accuracy and less errors.

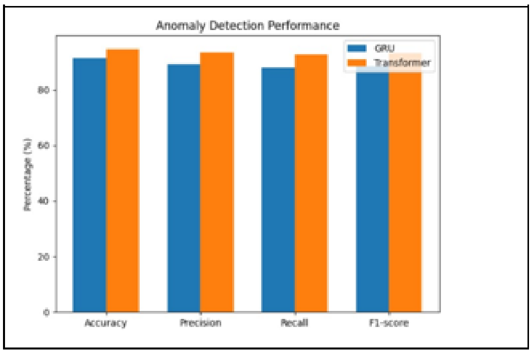

**Figure 10:** Metrics of GRU and Transformer

This bar graph presents the comparative anomaly detection performance of the GRU and Transformer models using standard evaluation metrics: Accuracy, Precision, Recall, and F1-score.

## 8. Conclusion

This work presents an effective deep learning–based framework for memory anomaly detection using a combination of GRU and Transformer models. The approach successfully captures both short-term and long-term temporal patterns in memory usage, enabling accurate and reliable anomaly detection. GRU provides stable predictions for normal behavior, while the Transformer effectively identifies sudden memory spikes. The combined model reduces false positives, improves sensitivity, and, with dynamic error-based thresholding, enhances detection reliability. Overall, the framework offers a scalable and automated solution for proactive system monitoring.

### Advantages

- Accurate detection of both normal patterns and abnormal memory spikes.
- Reduced false positives through model fusion
- Scalable and automated monitoring approach.

### Disadvantages

- Higher computational and resource requirements.
- Dependence on quality and quantity of training data.

### Declarations

a) Authors' Contributions
b) G. Lakshmi Supriya, contributed to the conceptualization of the study, literature review, and problem formulation. K. Mounika, was involved in methodology design, implementation of the GRU and Transformer models, and experimental analysis. Pearly Princess J, contributed to system architecture design, result analysis, manuscript preparation, and overall supervision of the study. *All authors read and approved the final manuscript.*

### Ethical Approval and Informed Consent

This study did not involve human participants, animals, or any personal or sensitive data. The work is based on system-generated and publicly accessible data. Therefore, ethical approval and informed consent were not required.

### Data Availability

The data used in this study were obtained from system monitoring logs and generated during experimental analysis. The datasets supporting the findings of this study are available from the corresponding author upon reasonable request.

## References

[1] May, P., Ehrlich, H. C., & Steinke, T. (2006, August). ZIB structure prediction pipeline: composing a complex biological workflow through web services. In *European Conference on Parallel Processing* (pp. 1148-1158). Berlin, Heidelberg: Springer Berlin Heidelberg.

[2] Czajkowski, K., Fitzgerald, S., Foster, I., & Kesselman, C. (2001, August). Grid information services for distributed resource sharing. In *Proceedings 10th IEEE International Symposium on High Performance Distributed Computing* (pp. 181-194). IEEE.

[3] Dip Das, J., Thulasiram, R. K., Henry, C., & Thavaneswaran, A. (2024). Encoder–decoder based LSTM and GRU architectures for stocks and cryptocurrency prediction. *Journal of Risk and Financial Management*, *17*(5), 200.

[4] Roy, K. S., & Bina, F. A. (2025). TweetGuard: Combining Transformer and Bi-LSTM Architectures for Fake News Detection in Large-Scale Tweets. *International Journal*, *11*(2), 23-45.

[5] ML, R., S, S., & R, N. (2025). An adaptive battery health monitoring framework using wavelet scattering and spiking graph transformers optimized by Arctic Wolf algorithm. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 1-27.

[6] Choudhary, A., & Arora, A. (2024). Assessment of bidirectional transformer encoder model and attention based bidirectional LSTM language models for fake news detection. *Journal of Retailing and Consumer Services*, *76*, 103545.

[7] Radhakrishnan, P., Ramar, V. A., Kushala, K., Induru, V., & SK, P. K. (2025). Online product decision support using review analysis and Transformer-GRU hybrid model through e-commerce platform. *Journal of Strategic Marketing*, 1-26.

[8] Annadurai, K., Saravanan, A., Kayalvili, S.,

Muniyandy, E., Aswani, I., & El-Ebiary, Y. A. B. (2025). Early Detection and Forecasting of Influenza Epidemics Using a Hybrid ARIMA-GRU Model. *International Journal of Advanced Computer Science & Applications*, *16*(5).

[9] Chhabra, A., & Alam, M. (2025). Reggru-opt: A robust gru-based rnn model for high-performance sentiment analysis and binary classification. *Arabian Journal for Science and Engineering*, 1-40.

[10] Sachan, R. C., & Malviya, R. K. (2024). Neural Transformers for Zero-Day Threat Detection in Real-Time Cybersecurity Network Traffic Analysis. *International Journal of Global Innovations and Solutions (IJGIS)*.

[11] Ma, M., Li, X., Fan, H., Qin, L., & Wei, L. (2025). Actual Truck Arrival Prediction at a Container Terminal with the Truck Appointment System Based on the Long Short-Term Memory and Transformer Model. *Journal of Marine Science and Engineering*, *13*(3), 405.

[12] Sihab-Us-Sakib, S., Rahman, M. R., Forhad, M. S. A., & Aziz, M. A. (2024). Cyberbullying detection of resource constrained language from social media using transformer-based approach. *Natural Language Processing Journal*, *9*, 100104.

[13] Hossain, M. M., Hossain, M. S., Hossain, M. S., Mridha, M. F., Safran, M., & Alfarhood, S. (2024). TransNet: deep attentional hybrid transformer for Arabic posts classification. *IEEE Access*.

[14] Burman, A., Solé-Casals, J., & Lew, S. E. (2024). Robust and memory-less median estimation for real-time spike detection. *Plos one*, *19*(11), e0308125.

[15] Boerlin, M., & Denève, S. (2011). Spike-based population coding and working memory. *PLoS computational biology*, *7*(2), e1001080.

[16] Rácz, M., Liber, C., Németh, E., Fiáth, R., Rokai, J., Harmati, I., ... & Márton, G. (2020). Spike detection and sorting with deep learning. *Journal of neural engineering*, *17*(1), 016038.

[17] Franke, F., Natora, M., Boucsein, C., Munk, M. H., & Obermayer, K. (2010). An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes. *Journal of computational neuroscience*, *29*(1), 127-148.

[18] Rodriguez-Perez, A., Ruiz-Amaya, J., Delgado-Restituto, M., & Rodriguez-Vazquez, A. (2012). A low-power programmable neural spike detection channel with embedded calibration and data compression. *IEEE transactions on biomedical circuits and systems*, *6*(2), 87-100.

[19] Lukito, V., Choi, E. J., Chang, I. J., Ha, S., & Je, M. (2025). A Spike Sorting SoC With Δ-Based Spike Detection and End-to-End Implementation of Autoencoder Feature Extraction Using Analog CIM. *IEEE Journal of Solid-State Circuits*.

[20] Medvedev, A. V., Agoureeva, G. I., & Murro, A. M. (2019). A long short-term memory neural network for the detection of epileptiform spikes and high frequency oscillations. *Scientific reports*, *9*(1), 19374.

[21] Prezioso, M., Mahmoodi, M. R., Bayat, F. M., Nili, H., Kim, H., Vincent, A., & Strukov, D. B. (2018). Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits. *Nature communications*, *9*(1), 5311.

[22] Li, J., Chen, X., & Li, Z. (2018). Spike detection and spike sorting with a hidden Markov model improves offline decoding of motor cortical recordings. *Journal of neural engineering*, *16*(1), 016014.

[23] Islam, M. S., Pourmajidi, W., Zhang, L., Steinbacher, J., Erwin, T., & Miranskyy, A. (2021, May). Anomaly detection in a large-scale cloud platform. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 150-159). IEEE.

[24] Chen, J., Chakraborty, J., Clark, P., Haverlock, K., Cherian, S., & Menzies, T. (2019, August). Predicting breakdowns in cloud services (with SPIKE). In *Proceedings of the 2019 27th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering* (pp. 916-924).

[25] Meyer, L. M., Zamani, M., Rokai, J., & Demosthenous, A. (2024). Deep learning-based spike sorting: a survey. *Journal of Neural Engineering*, *21*(6), 061003.