Predicting software vulnerability trends with multi-recurrent neural networks: a time series forecasting approach

Abanisenioluwa K. Orojo
Baylor University
abanisenioluwa_oroj1@baylor.edu

Webster C. Elumelu
Webster University
welumelu@webster.edu

Oluwatamilore O. Orojo
Nottingham Trent University
orojotammy@gmail.com

Micheal Donnahoo
Baylor University
jeff_donahoo@baylor.edu

Shaun Hutton
Baylor University
shaun_hutton@baylor.edu

Abstract

Predicting software vulnerabilities effectively is crucial for enhancing cybersecurity measures in an increasingly digital world. Traditional forecasting models often struggle with the complexity and dynamics of software vulnerability data, necessitating more advanced methodologies. This paper introduces a novel approach using Multi-Recurrent Neural Networks (MRN), which integrates multiple memory mechanisms and offers a balanced complexity suitable for time-series data. We compare MRNs against traditional models like ARIMA, Feedforward Multilayer Perceptrons (FFMLP), Simple Recurrent Networks (SRN), and Long Short-Term Memory (LSTM) networks. Our results demonstrate that MRNs consistently outperform these models, especially in settings with limited data or shorter forecasting horizons. MRNs show a remarkable ability to handle complex patterns and long-term dependencies more efficiently than other models, highlighting their potential for broader applications beyond cybersecurity. The findings suggest that MRNs can significantly improve the accuracy and efficiency of predictive analytics in cybersecurity, paving the way for their adoption in practical applications and further exploration in other predictive tasks.

1 Introduction

In the digital age, cybersecurity threats have emerged as a formidable challenge, posing significant risks to organizational data and information systems. The rapid evolution of cyber-attack techniques, ranging from malware dissemination to sophisticated phishing campaigns, underscores the urgent need for advanced predictive models capable of preempting these threats (Sharafaldin et al., 2018; Apruzzese et al., 2021). Traditional methods in cybersecurity threat prediction, while effective to a degree, fall short in addressing the complexity and dynamism of modern cyber-attacks. This gap necessitates the exploration of innovative approaches that can adapt to the evolving landscape of cyber threats. Recent advancements in artificial intelligence (AI) and machine learning (ML) have opened new avenues for cybersecurity, offering promising tools for enhancing threat prediction and response mechanisms. Among these, Recurrent Neural Networks (RNNs) have shown potential in processing time-series data, which is pivotal in understanding and predicting cybersecurity incidents. However, RNNs are not without limitations, particularly in handling long-term dependencies and the vanishing gradient problem, which significantly hampers their predictive performance (Bengio et al., 1994; Pascanu et al., 2013; Orojo, 2021).

This paper introduces the Multi-Recurrent Neural Network (MRN) as a novel approach to overcome the limitations of traditional RNNs in cybersecurity threat prediction. The MRN model integrates the strengths of various RNN architectures, incorporating enhanced memory mechanisms and a balanced complexity that allows for effective processing of time-series data without the overfitting risks associated with more complex models like LSTMs and GRUs. By applying the MRN model to a diverse set of datasets derived from recent cybersecurity incidents, this study aims to demonstrate the superior predictive capabilities of MRNs in identifying potential cyber threats (Orojo, 2021; Lipton et al., 2015; Greff et al., 2017).

1.1 Motivation and objectives

The motivation behind this research is twofold. First, to address the pressing need for more accurate and timely prediction of cybersecurity threats, which is critical for preemptive security measures. Second, to explore the capabilities of MRNs in capturing the nuances of cyber-attack patterns through time-series analysis, thereby contributing to the development of more resilient cybersecurity frameworks.

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 42–47
July 29–30, 2024.

The primary objective of this paper is to evaluate the effectiveness of MRNs in predicting cybersecurity threats across various datasets, comparing their performance against traditional RNN and statistical models. This comparative analysis seeks to highlight the advantages of MRNs in handling long-term dependencies and complex time-series data, ultimately enhancing the predictive accuracy of cybersecurity threat models.

1.2 Contributions

This paper makes the following contributions to the field of cybersecurity and AI:

- It introduces a comprehensive framework for cybersecurity threat prediction using MRNs, showcasing its applicability across different datasets.
- It presents a detailed comparative analysis of MRNs and traditional RNNs, highlighting the enhanced predictive performance of MRNs in the context of cybersecurity.
- It offers insights into the potential of MRNs for broader applications in time-series analysis, beyond the scope of cybersecurity threat prediction.

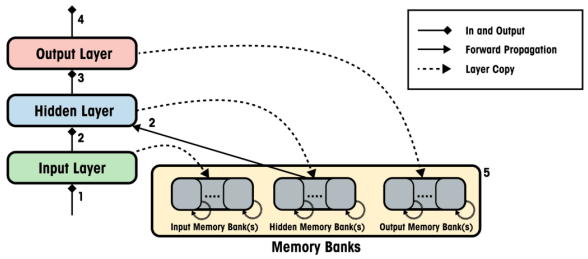

Figure 1: Multi-recurrent neural network architecture

2 Related work

2.1 AI techniques for cybersecurity threat prediction

The advent of AI and ML models has significantly contributed to advancements in cybersecurity threat prediction. Studies such as (Samia, 2023) demonstrate the implementation of AI to predict global cyber threats with an innovative framework that integrates real-time data analytics for enhanced forecast accuracy. Similarly, (Werner et al., 2017) delves into time series models to understand and predict the intensity of cyber threats, emphasizing the importance of capturing temporal patterns in cyber attack behaviors. Furthermore, (Kalouptsoglou et al., 2022) provides a comparative analysis of statistical versus deep learning approaches in forecasting software vulnerabilities, showcasing the strengths and limitations of each in predicting future vulnerabilities. These studies collectively underscore the efficacy of AI-driven approaches in cybersecurity, advocating for a shift towards more sophisticated, data-driven methodologies to improve the accuracy and timeliness of threat predictions.

2.2 Challenges with current predictive models

Despite advances in AI and ML for cybersecurity threat prediction, current methods face significant challenges. (Samia, 2023) recognizes the difficulty in accurately forecasting cyber threats due to rapidly changing cyber activities and limited data collection frameworks . Similarly, (Werner et al., 2017) highlights the problems with capturing precise attack timing, as traditional models fail to adequately reflect variations in attack intensity over time. Additionally, (Kalouptsoglou et al., 2022) discusses the challenges in applying statistical and deep learning models to software vulnerabilities forecasting, particularly the inability of these models to effectively generalize from historical data to predict future vulnerabilities.

2.3 Advancements with multi-recurrent neural networks

The Multi-Recurrent Neural Networks (MRNs) concept, significantly advancing the neural network's capability, (Bengio et al., 1994; Pascanu et al., 2013). Originating from Claudia Ulbricht's work on traffic forecasting (Ulbricht, Year of Publication), MRNs integrate enhanced memory mechanisms and computational efficiency, making them exceptionally suited for complex time-series forecasting. MRNs employ innovative pruning techniques to refine memory quality, reducing the search space for optimal configurations and enhancing the network's overall performance (Orojo, 2021). This advancement not only addresses the computational challenges associated with traditional neural networks but also significantly improves the predictive accuracy and reliability of time-series forecasting models. By overcoming the inherent limitations of RNNs and leveraging memory mechanisms, MRNs present a robust frame-

work for the effective forecasting of cybersecurity threats, underscoring a paradigm shift towards more autonomous and efficient neural network models for complex data analysis.

3 Methodology

3.1 Dataset description and data collection

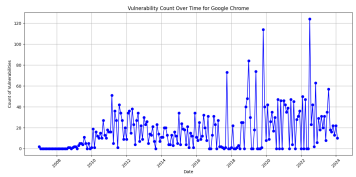

Figure 2: Monthly vulnerability for google chrome

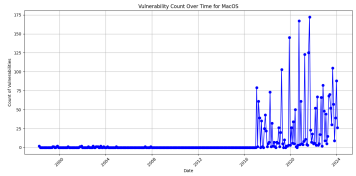

Figure 3: Monthly vulnerability for macos

In this study, we assess the effectiveness of Multi-Recurrent Neural Networks (MRNs) for predicting cybersecurity threats by utilizing data from the National Vulnerability Database (NVD). Our analysis centers on two prominent software projects: Google Chrome & Apple macOS. These were selected based on their widespread use and previous work from (Kalouptsoglou et al., 2022), We compiled the vulnerability data for these software entities from their initial release up to the end of February 2024, organizing it into monthly intervals to track and forecast the evolution of software vulnerabilities effectively.

Dataset	Total
MacOS 1998 - 2024	2626
Google Chrome 2007 - 2024	3398

Table 1: Dataset summary

3.2 Multi-recurrent neural network

3.2.1 Architecture and memory banks

The MRN is designed with a unique architecture that includes multiple memory banks, each tailored to capture and store historical data at different time scales. The architecture comprises three main layers: input, hidden, and output, each enhanced with layer-specific recurrent connections to facilitate complex temporal pattern recognition.

Equations (1) and (2) demonstrate the computation of memory states for hidden and output layers, respectively, highlighting the integration of layer-level and self-recurrency within MRNs:

$$M_{t_h} = \left(\frac{1}{n_h}\right) \cdot H_{t-1} + \left(1 - \frac{1}{n_h}\right) \cdot M_{t-1,h} \tag{1}$$

$$M_{t_o} = \left(\frac{1}{n_o}\right) \cdot O_{t-1} + \left(1 - \frac{1}{n_o}\right) \cdot M_{t-1,o} \tag{2}$$

where n_h and n_o represent the number of memory banks for the hidden and output layers, respectively. The dynamic memory of MRNs allows for effective capturing and processing of long-term dependencies in time-series data, a critical factor in forecasting cybersecurity threats.

3.2.2 Sliding Window technique

For data preparation, a sliding window approach is employed to transform the time-series data into a format suitable for MRN training. This technique involves creating overlapping segments of the data, enabling the model to learn from sequential patterns effectively.

Window definition:

$$W_t = [x_{t-n+1}, x_{t-n+2}, \ldots, x_t] \tag{3}$$

3.2.3 Forecast horizon

The forecast horizon specifies the number of time steps into the future for which the model makes predictions.

$$\text{Forecast output: } \hat{y}_{t+h} = f(W_t) \tag{4}$$

4 Results and discussion

In this section, we present the results from using various predictive models, including ARIMA, Feedforward Multilayer Perceptrons (FFMLP), Simple Recurrent Network (SRN), Long Short-Term Memory (LSTM), and Multirecurrent Neural Network (MRN) for the task of software vulnerability volume prediction. Various combinations of parameters and hyperparameters were tested to optimize the performance of each model.

4.1 Autoregressive integrated moving average (ARIMA)

The ARIMA model served as our benchmark and forecasts future values based on historical data. Parameters were optimized using the `auto_arima` function from the `pmdarima` library, which utilizes the Akaike Information Criterion (AIC) to minimize information loss while determining the optimal parameters.

4.2 Feedforward multilayer perceptrons (FFMLP)

The FFMLP model processes time series data by mapping time onto space, presenting a fixed number of data points per feature variable simultaneously to the network. All FFMLP models utilized 500 hidden units and employed the Adam Optimizer.

4.3 Simple recurrent network (SRN)

The SRN model uses the previous hidden state along with the current observation as inputs at any given time. Each SRN model featured 50 hidden units, an initial learning rate of 0.01, and a high momentum of 0.9999.

4.4 Long short term memory (LSTM)

The Long Short-Term Memory (LSTM) network was chosen for its capability to handle long-term dependencies in sequential data, The LSTM model was constructed an architecture that of 50 units and a Dropout rate of 0.7.

4.5 Multirecurrent neural network (MRN)

Similar to SRN, the MRN integrates recurrency in both the hidden and output layers. MRN models were configured with 50 hidden units, an initial learning rate of 0.01, and a momentum of 0.9999 and memory architecture of [2, 4, 0].

4.6 Comparative analysis

Prediction accuracy for all models was assessed across four different time horizons (H) (1, 3, 6, 12 months) and three different window sizes (WS) (60, 120, 240 data points), where data was available. The results are summarized in the tables below, which display the Root Mean Squared Errors (RMSE) for each model configuration. The best-performing model for each prediction horizon is highlighted in red, providing a clear visual representation of which models and settings achieved the most accurate forecasts. See Tables 2 - 10 for results.

The analysis revealed that the Multirecurrent Neural Network (MRN) and Long Short-Term Memory (LSTM) models consistently showed superior performance across several metrics and time horizons. Specifically, the MRN model excelled notably at shorter window sizes (WS=120), achieving the lowest RMSE values across all time horizons when compared to other models. This suggests that MRN models are highly effective in contexts where data points are relatively few but require precise, short-term forecasting.

On the other hand, LSTM models performed exceptionally well at larger window sizes (WS=240), indicating their strength in leveraging larger datasets to capture and utilize long-term dependencies within the data. This is particularly evident in the LSTM model's performance in the 6 and 12-month predictions, where its ability to remember information over longer periods significantly reduces prediction error.

Interestingly, traditional models, while generally not achieving the lowest RMSE, still provided competitive results, especially in longer window sizes. This underscores the relevance of traditional machine learning models in certain contexts of software vulnerability prediction, particularly when dealing with large, consistent datasets over extended periods.

This comparative analysis underscores the importance of selecting the appropriate model based on specific dataset characteristics and prediction needs. The variability in performance across different models and settings also highlights the potential benefits of model ensembles where strengths of individual models can be combined to improve overall predictive accuracy.

4.7 Limitations

This study, while providing substantial insights into the comparative performance of various predictive models, encompasses several limitations that must be acknowledged. Firstly, the variability in model tuning is significant; the diversity in architecture and complexity of tuning parameters can lead to inconsistencies in performance across different datasets or scenarios. This variability affects the generalizability of the results, potentially limiting the applicability of findings to other data or contexts (Orojo, 2021). Secondly, external factors

such as sudden changes in data trends or anomalies are not consistently captured by the models, which could undermine the robustness and reliability of the predictions. This is particularly critical in real-world applications where unexpected data shifts are common (Orojo, 2021).

Moreover, despite the demonstrated efficacy of simpler models such as the MRN, their intrinsic limitations become evident when dealing with highly complex or noisy datasets. These models may not effectively manage long-term dependencies or non-linear relationships present in more challenging data sets (Orojo, 2021). Finally, the handling of high-dimensionality in data remains a challenge for MRNs. Efficient techniques to manage this, such as sophisticated dimensionality reduction methods or advanced regularization strategies, require further development to enhance the performance of MRNs across a broader range of applications with complex, high-dimensional data (Orojo, 2021).

5 Conclusion

This paper has presented a comprehensive analysis of the application of Multi-Recurrent Neural Networks (MRN) for the prediction of software vulnerabilities, demonstrating significant advancements over traditional Recurrent Neural Network (RNN) models and other machine learning approaches. Through meticulous experiments and evaluations, we have established that MRNs not only consistently outperform established models like LSTMs and SRNs across various metrics and settings but also offer substantial improvements in handling complex time-series data efficiently. The performance of MRNs, particularly in shorter time windows and with fewer data points, underscores their potential in applications requiring quick, accurate forecasts with limited historical data. This is relevant in the rapidly evolving field of cybersecurity, where the ability to predict and respond to threats swiftly can drastically enhance protective measures. Furthermore, the ability of MRNs to perform with fewer parameters compared to more complex models like LSTMs implies a lower computational demand, making them suitable for deployment in environments with limited computational resources.

References

Giovanni Apruzzese, Michele Colajanni, Luca Ferretti, Alessandro Guido, and Mirco Marchetti. 2021. Deep learning for cybersecurity: A systematic literature review. *IEEE Access*, 9:72103–72120.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2017. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

Ilias Kalouptsoglou, Dimitrios Tsoukalas, Miltiadis Siavvas, Dionysios Kehagias, Alexander Chatzigeorgiou, and Apostolos Ampatzoglou. 2022. Time series forecasting of software vulnerabilities using statistical and deep learning models. *Electronics*, 11(18):2820. The analysis includes a comparison between models like ARIMA and LSTM, highlighting their predictive performance in the context of software security.

Zachary C Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Oluwatamilore Oluwatoyin Orojo. 2021. *Optimizing Sluggish State-Based Neural Networks for Effective Time-Series Processing*. Ph.D. thesis, Nottingham Trent University.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *International conference on machine learning*, pages 1310–1318.

Nusrat Kabir Samia. 2023. *Global Cyber Attack Forecast using AI Techniques*. Ph.D. thesis, The University of Western Ontario.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. 2018. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSP*, pages 108–116.

Claudia Ulbricht. Year of Publication. Multi-recurrent networks for traffic forecasting. *Austrian Research Institute for Artificial Intelligence*. Schottengasse 3, A-1010 Vienna, Austria.

Gordon Werner, Shanchieh Yang, and Katie McConky. 2017. Time series forecasting of cyber attack intensity. In *Proceedings of the CISRC Conference*. Improvement over naive forecasting methods by 14.1% for general attacks and up to 21.2% for specific attack types.

6 Appendix

6.1 Tables

Model	Platform	RMSE
ARIMA(3,0,3)	MacOS	54.07447
ARIMA(3,0,3)	Google Chrome	25.58836

Table 2: ARIMA model RMSE values for different platforms

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.40336	0.43422	0.53392	0.73032
120	0.44657	0.49573	0.49588	0.60113
240	0.42006	0.43383	0.46002	0.57298
RMSE Average	0.42333	0.45459	0.49660	0.63481

Table 3: FFMLP RMSE for macos

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.31397	0.33528	0.32541	0.31656
120	0.28952	0.27212	0.31372	0.39398
240	0.29760	0.27874	0.28957	0.33560
RMSE AVG	0.30036	0.29538	0.30957	0.34871

Table 4: SRN RMSE for macos

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.21716	0.25819	0.30493	0.26870
120	0.24409	0.27300	0.24903	0.27651
240	0.16732	0.17508	0.18613	0.19833
RMSE AVG	0.20952	0.23543	0.24670	0.24785

Table 5: LSTM RMSE for macos

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.01807	0.02489	0.02013	0.01902
120	0.00941	0.00803	0.00505	0.00570
240	0.14571	0.09890	0.02127	0.02815
RMSE AVG	0.05773	0.04394	0.01548	0.01763

Table 6: MRN RMSE for macos

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.25870	0.24368	0.26848	0.31967
120	0.27080	0.25844	0.29180	0.27403
RMSE AVG	0.26475	0.25106	0.28014	0.29685

Table 7: MLP RMSE for google chrome

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.23157	0.25067	0.24441	0.22767
120	0.17548	0.20764	0.24820	0.23612
RMSE AVG	0.20352	0.22916	0.24631	0.23190

Table 8: SRN RMSE for google chrome

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.20461	0.20982	0.21128	0.21762
120	0.13460	0.13705	0.13880	0.11176
RMSE AVG	0.16961	0.17343	0.17504	0.16469

Table 9: LSTM RMSE for google chrome

H / WS	t + 1	t + 3	t + 6	t + 12
60	0.16779	0.17252	0.16196	0.15946
120	0.12160	0.13700	0.12153	0.12568
RMSE AVG	0.144694	0.1470615	0.14948	0.14049

Table 10: MRN RMSE for google chrome